Zur Theorie und Anwendung von Näherungsverfahren für die Lösung nichtlinearer Operatorgleichungen

Helmut Maier

1972

ZUR THEORIE UND ANWENDUNG VON NÄHERUNGSVERFAHREN FÜR DIE LÖSUNG NICHTLINEARER OPERATORGLEICHUNGEN

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der Fakultät für Mathematik
der Ludwig-Maximilians-Universität zu München

vorgelegt von

Helmut Maier
aus Singen

München 1972

UNIV.
LEIDEN
BIBL.

1. Berichterstatter: Prof. Dr. G. Hämmerlin

2. Berichterstatter: Prof. Dr. E. Wienholtz

Tag der mündlichen Prüfung: 14.11.1972

Inhaltsverzeichnis

	Seite
EINLEITUNG	1
ERSTES KAPITEL	
DISKRETE KONVERGENZ KOMPAKTER NICHTLINEARER OPERATOREN	6
§ 1. Begriffe, Problemstellung	6
§ 2. Ein Konvergenzsatz für die diskrete Approximation eines Operators durch eine diskret kompakte Folge	10
§ 3. Konvergenz von Eigenwerten, Eigenvektoren und stetigen Zweigen	14
ZWEITES KAPITEL	
KONVERGENZ DIFFERENZIERBARER OPERATOREN UND NEWTON-VERFAHREN	22
§ 4. Ein Konvergenzsatz für die Approximation von Gleichungen der Form $Kx = 0$	22
§ 5. Der Konvergenzsatz für Gleichungen zweiter Art: $x = Qx + f$	30
§ 6. Der Konvergenzsatz für Hammersteinsche Gleichungen $x = LNx + f$	35
DRITTES KAPITEL	
EIN VERFAHREN ZUR PRAKTISCHEN BESTIMMUNG DER LÖSUNG EINER HAMMERSTEINSCHEN INTEGRALGLEICHUNG	41
§ 7. Definition und Konvergenz des Ersatzsystems	41
§ 8. Abschätzung des Diskretisierungsfehlers	48
§ 9. Berechnung der Gewichte bei verallgemeinerter Quadratur	55
§ 10. Erfassung der Rundefehler durch Methoden der Intervallarithmetik	59
§ 11. Numerische Beispiele	64
LITERATURVERZEICHNIS	72
LEBENSLAUF	

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41546014_0026

Einleitung

Viele Aufgaben der nichtlinearen Analysis lassen sich als Gleichung der Form $Kx = 0$ schreiben; hierbei ist x ein Element aus einem geeigneten Funktionenraum E und K ein nichtlinearer Operator, etwa ein Differential-, Integral- oder sonstiger Funktionaloperator. Für eine große Klasse von Problemen ist es zweckmäßig und ausreichend, als Funktionenraum einen Banachraum zu wählen. Bei der Untersuchung von Existenz- und Eindeutigkeitsfragen hat es sich gezeigt, daß in erster Linie Eigenschaften des Operators K eine Rolle spielen, die gegenüber "kleinen" Änderungen der Gleichung invariant sind. Eine der wichtigsten topologischen Invarianten ist der Abbildungsgrad.

Da die Theorie der linearen Operatoren natürlich schon länger untersucht und weiter vorangeschritten war, lag es bei der Entwicklung der nichtlinearen Analysis in Funktionenräumen nahe, wie in der reellen Analysis nichtlineare Probleme wenigstens lokal zu linearisieren, um dann die vorhandene lineare Theorie anwenden zu können. Dies führte zu dem wesentlichen Begriff der Ableitung eines Operators.

In der angewandten Mathematik tritt zur Frage der Existenz und Eindeutigkeit fast immer das Problem der praktischen Lösung einer Gleichung. Die Komplexität nichtlinearer Aufgaben macht hierzu im allgemeinen die Verwendung von Rechenanlagen unumgänglich. Endliche Rechenzeit und endliche Menge der Maschinenzahlen machen - außer etwa bei Analogrechnern - eine Diskretisierung des meistens kontinuierlichen Problems erforderlich; das bedeutet den Übergang zu einer Näherungsgleichung $\tilde{K}\tilde{x} = 0$. Wenn es möglich ist, den Näherungsoperator so zu wählen, daß etwa der Abbildungsgrad erhalten bleibt oder das linearisierte Problem "ähnlich" ausfällt, kann unter geeigneten Voraussetzungen auch die Lösung $\tilde{x}^*$ von $\tilde{K}\tilde{x} = 0$ (falls vorhanden) nahe bei der exakten Lösung x^* von $Kx = 0$ liegen.

KRASNOSEL'SKII [1964] hat Gleichungen der Form $Kx = x$ mit vollstetigem Operator K untersucht. Bei Integraloperatoren kann man in vielen Fällen unter nicht allzu einschränkenden Voraussetzungen die Vollstetigkeit des Operators realisieren. Daher ist es beispielsweise oft zweckmäßig, Differentialgleichungen in Integralgleichungen umzuformen. Es gibt natürlich auch andere Methoden, um zu vollstetigen Operatoren zu gelangen, etwa die Sobolewschen Einbettungssätze bei elliptischen Gleichungen.

Die übliche Störungstheorie für nichtlineare Gleichungen der Form $Kx = x$ betrachtet nun Folgen von Operatoren $K_i : E \to E$ $(i \in \mathbb{N})$, die in geeigneter Weise gegen K konvergieren (z.B. VAINIKKO [1966]). KRASNOSEL'SKII [1964], S. 170 untersucht den Fall $K_i = P_i K$, wobei P_i eine Projektion auf einen endlichdimensionalen Teilraum ist. Für differenzierbare Operatoren gibt es Ergebnisse von AUBIN [1967], TUCKER [1969], VAINIKKO [1969]. Die zuerst genannte Arbeit war für STUMMEL [1970] ein Ausgangspunkt für eine Verallgemeinerung der linearen Störungstheorie auf diskret konvergente Operatorfolgen $K_i : E_i \to E_i$, wobei die Räume E_i den Raum E möglichst allgemein approximieren. Ausgehend von der dortigen Begriffsbildung im linearen Fall wird in der vorliegenden Arbeit ein Konvergenzsatz für nichtlineare Operatoren bewiesen.

Der Hauptsatz (2.1) von Kapitel I zeigt, daß sich die Fixpunkteigenschaft eines vollstetigen Operators K auf eine diskret kompakte, diskret konvergente Folge vollstetiger Operatoren K_i überträgt, falls die abgeschlossenen Teilräume E_i eine spezielle diskrete Approximation von E bilden. Die wesentlichen Hilfsmittel zum Beweis dieses Satzes sind der Leray-Schaudersche Projektionsoperator und die Drehungsinvarianz homotoper Vektorfelder. In § 3 erhalten wir, als Folgerung aus dem Hauptsatz, die Konvergenz der Eigenwerte und Eigenvektoren der Operatoren K_i gegen Eigenwert und Eigenvektor von K. Mit Hilfe einer leichten Verallgemeinerung

des Hauptsatzes können wir auch die Konvergenz stetiger Zweige zeigen. Ein Anwendungsbeispiel ist die Konvergenz der gewöhnlichen Kollokation bei nichtlinearen Integralgleichungen. Fehlerabschätzungen aus dieser Theorie scheinen kaum möglich.

Eine Möglichkeit der Linearisierung eines nichtlinearen Operators bietet das Konzept der Fréchetableitung. Eine für die Anwendung bedeutungsvolle Theorie wurde von KANTOROWITSCH (KANTOROWITSCH-AKILOW [1964], Kap. XVIII) dargestellt. Auf dieser Grundlage entwickeln wir in Kapitel II eine Störungstheorie des Newton-Verfahrens und beweisen einen Konvergenzsatz. Ähnliche Ergebnisse wie in § 4 erhielt MOORE [1968] unter der Voraussetzung der kollektiven Kompaktheit - ein Spezialfall der diskreten Kompaktheit in Kapitel I - der Folgen $\{K_i\}$ und $\{K_i'(x)\}$ $(i \in \mathbb{N})$ und der Kompaktheit von K und $K'(x)$ (x ist irgendein Element einer geeigneten offenen Umgebung D der Lösung). Wir können auf diese Voraussetzungen verzichten zugunsten der Konvergenzforderung $\|K_i'(x_0) - K'(x_0)\| \to 0$ (für ein festes x_0); MOORE benötigt die Konvergenz $\|K_i'(x)h - K'(x)h\| \to 0$ (für $(x,h) \in D \times E$). Hauptsatz (4.4) zeigt, daß sich die Eigenschaft der Gleichung $Kx = 0$ mit dem Newton-Verfahren lösbar zu sein unter den Konvergenzforderungen

$$\begin{aligned} &\|K_i x - Kx\| \to 0 && \text{in } \{x : \|x-x_0\| \le r\} \\ &\|K_i'(x_0)-K'(x_0)\| \to 0 && \\ &\|K_i''(x)-K''(x)\| \to 0 && \text{gleichmäßig in } \{x : \|x-x_0\| \le r\} \end{aligned}$$

auf $K_i x_i = 0$ überträgt (und umgekehrt).

§ 5 und § 6 bringen eine Anwendung auf Hammersteinsche Operatoren $Kx = x - LNx - f$ ($f \in E$ fest). Es zeigt sich, daß eine getrennte Diskretisierung des linearen Anteils L und des nichtlinearen Anteils N des Hammersteinoperators möglich - und wie sich in Kapitel III zeigt - auch vorteilhaft ist.

Eine Anwendung der Theorie des zweiten Kapitels auf Hammersteinsche Integralgleichungen der Form

$$x(s) = \int_0^1 L(s,t)N(t,x(t))dt + f(s)$$

stellen wir in Kapitel III dar. Durch Lagrange-Interpolation bezüglich der für numerische Zwecke besonders günstigen Tschebyscheffknoten (T-Knoten) $s_k^{(n)} = \cos\frac{k\pi}{n}$ $(k=0,1,\ldots,n)$ gewinnen wir das Ersatzsystem für die Unbekannten $a_i^{(n)}$

$$a_i^{(n)} - \sum_{k=0}^{n} A_{ik}\, N(s_i^{(n)}, a_k^{(n)}) - f(s_i^{(n)}) = 0 \quad (i=0,\ldots,n),$$

wobei

$$A_{ik} = \int_0^1 L(s_i^{(n)},t)l_k^{(n)}(t)dt$$

und $l_k^{(n)}(t)$ das k-te Lagrange-Grundpolynom vom Grad n bedeutet. Zu einer Abschätzung des Diskretisierungsfehlers gelangen wir in § 8, indem wir ein Ergebnis von MYSOVSKIH [1962] zur Lösung von Integralgleichungen durch Quadraturkollokation auf Kollokation mit verallgemeinerter Quadratur übertragen. Hierbei erweisen sich Fehlerabschätzungen bei der Clenshaw-Curtis-Quadratur, die von FRASER-WILSON [1966] und LOCHER [1969] angegeben wurden, auch bei verallgemeinerter Quadratur als brauchbar.

Eine Möglichkeit, die Gewichte A_{ik} näherungsweise durch rekursiv schnell berechenbare Summen zu bestimmen, geben wir in § 9 an. Für den wichtigen Spezialfall, daß $L(s,t)$ eine Greensche Funktion ist, lassen sich die Gewichte besonders einfach exakt angeben.

Im allgemeinen Fall werden die Größen A_{ik} aber mit Fehlern behaftet sein, die wir außer den Rundefehlern beim Lösen des Ersatzsystems durch das Newton-Verfahren auf einer Rechen-

anlage zu berücksichtigen haben. Wie wir hierzu Ergebnisse von ROKNE-LANCASTER [1969] und ALEFELD-HERZBERGER [1970] verwenden, ist der Inhalt des § 10. Die numerischen Beispiele des § 11 werden mit der gewöhnlichen Kollokation verglichen, wobei sich eine gewisse Überlegenheit unseres Verfahrens besonders für kleines n und die Brauchbarkeit der Fehlerschranken zeigt.

An dieser Stelle möchte ich meinem akademischen Lehrer, Herrn Prof. Dr. G. Hämmerlin, für seine stets ermutigende Unterstützung beim Entstehen dieser Arbeit herzlich danken.

Mein besonderer Dank gilt auch Frl. stud.math. H. Damek für ihre sehr gewissenhafte Hilfe bei der Programmierung und Durchführung der numerischen Beispiele.

ERSTES KAPITEL

DISKRETE KONVERGENZ KOMPAKTER NICHTLINEARER OPERATOREN

§ 1. Begriffe, Problemstellung

Seien E_i, $i \in \mathbb{N}_o$, lineare Teilräume eines reellen Banachraums E. STUMMEL [1970], S. 85, gibt eine notwendige und hinreichende Bedingung dafür an, daß diese Teilräume eine "diskrete Approximation" bilden. Der Einfachheit halber wollen wir hier den Begriff der diskreten Approximation durch diese Bedingung definieren.

(1.1) Definition

Die Teilräume E_i, $i \in \mathbb{N}$, bilden eine diskrete Approximation des Teilraumes E_o genau dann, wenn

$$(1.1.1) \qquad \bigwedge_{u \in E_o} |u, E_i| \xrightarrow[i \to \infty]{} 0, \text{ wobei } |u, E_i| := \inf_{w \in E_i} \|u - w\|.$$

(1.2) Bemerkungen

(1.2.1) Hinreichend für Bedingung (1.1.1) ist die Existenz einer dichten Teilmenge $\Phi \subset E_o$, so daß zu jedem $\varphi \in \Phi$ eine Zahl $i(\varphi) \in \mathbb{N}$ mit der Eigenschaft

$$\bigwedge_{i \geq i(\varphi)} \varphi \in E_i \quad \text{existiert.}$$

(1.2.2) Im Falle

$$E_1 \subset E_2 \subset \ldots \subset \bigcup_{i \in \mathbb{N}} E_i =: \Phi$$

ist die Bedingung in (1.2.1) erfüllt.

Beweis von (1.2.1):

Zu $u \in E_o$ und $\varepsilon > 0$ beliebig gibt es $\tilde{u} \in \tilde{E}$ mit $\|u-\tilde{u}\| \leq \varepsilon$. Zu $\tilde{u}(\varepsilon)$ gibt es ein $i(\varepsilon)$ mit $\tilde{u} \in E_i$, $i \geq i(\varepsilon)$. Also ist für $i \geq i(\varepsilon)$

$$|u,E_i| \leq \varepsilon + \inf_{w \in E_i} \|\tilde{u}-w\| = \varepsilon .$$

Im folgenden sei stets vorausgesetzt, daß die Folge $\{E_i\}_{i \in \mathbb{N}}$ eine diskrete Approximation von E_o bildet.

Nun wollen wir den Begriff einer diskret kompakten Folge stetiger Operatoren definieren.

(1.3) Definition

Eine Folge stetiger linearer Operatoren $K_i : E_i \to E_i$, $i \in \mathbb{N}$ heißt diskret kompakt, wenn es für jede Teilfolge $I' \subset \mathbb{N}$ und jede beschränkte Folge $\{u_i\}_{i \in I'}$ eine Teilfolge $I'' \subset I'$ und ein Element $w \in E_o$ gibt, mit der Eigenschaft

$$\|K_i u_i - w\| \to 0 \quad (i \in I'' \subset I' \subset \mathbb{N}) .$$

Ferner benötigen wir den Begriff der diskreten Konvergenz von Operatorfolgen.

(1.4) Definition

Die Folge stetiger linearer Operatoren $K_i : E_i \to E_i$, $i \in \mathbb{N}$ heißt diskret konvergent gegen den stetigen linearen Operator $K : E_o \to E_o$, wenn für jedes $u \in E_o$ und jede konvergente Folge $u_i \in E_i$, $i \in \mathbb{N}$ gilt:

$$u_i \to u \Rightarrow K_i u_i \to Ku .$$

Eng verbunden mit der diskreten Konvergenz von Operatoren sind die bekannten Begriffe der Stabilität und Konsistenz.

(1.5) Definition

Die Folge stetiger linearer Operatoren $K_i : E_i \to E_i$, $i \in \mathbb{N}$ heißt konsistent bzgl. $K : E_o \to E_o$, wenn es eine dichte Teilmenge $D \subset E_o$ gibt und zu jedem $d \in D$ eine Folge von Elementen $d_i \in E_i$, $i \in \mathbb{N}$ mit der Eigenschaft

$$\|d_i - d\| \to 0 \Rightarrow \|K_i d_i - Kd\| \to 0 .$$

Eine Definition der Stabilität für nichtlineare Operatoren findet man bei AUBIN [1967]. Da wir diese nicht benötigen werden, sei hier nur die einfachere Definition im Falle linearer Operatoren aufgeführt.

(1.6) Definition

Eine Folge linearer Operatoren $K_i : E_i \to E_i$, $i \in \mathbb{N}$ heißt stabil, wenn

$$\bigvee_{a \in \mathbb{R}} \bigwedge_{i \in \mathbb{N}} \|K_i\| \leq a.$$

Im Falle linearer Operatoren besteht nun ein Zusammenhang zwischen den Begriffen der letzten drei Definitionen (s. STUMMEL [1970], S. 53).

(1.7) SATZ

Eine Folge stetiger linearer Operatoren ist genau dann diskret konvergent, wenn sie stabil und konsistent ist.

Bei nichtlinearen Operatoren ist eine solche Charakterisierung der diskreten Konvergenz nicht möglich. Es zeigt sich nämlich, daß die Konsistenz und die Stabilität (letztere im Sinne von AUBIN) lediglich hinreichend für die diskrete Konvergenz sind. Bei den von uns angestrebten Aussagen werden wir stets die diskrete Konvergenz der betrachteten Operatorfolgen voraussetzen.

Wir wollen uns nun der Frage zuwenden, wie man Fixpunktgleichungen

$$Kx = x$$

für nichtlineare Operatoren diskret approximieren kann. Hierzu benötigen wir zur Charakterisierung der Operatoren die Theorie von KRASNOSEL'SKII [1964], insbesondere die Begriffe der Rotation eines Vektorfeldes und des Abbildungsgrades. Da es sich um lokale Aussagen handeln wird, nehmen wir an, daß $B \subset E_0$ ein offenes, beschränktes Gebiet ist. Der Operator K sei auf $\overline{B} := B \cup \partial B$ definiert:

$$K : \overline{B} \to E_0 .$$

Die Menge $B_i := B \cap E_i$ ist offen in E_i. $\overline{B}_i := B_i \cup \partial B_i$ ist der Abschluß in E_i. Wir nehmen an, daß die Operatoren K_i auf $\overline{B}_i$ definiert sind:

$$K_i : \overline{B}_i \to E_i .$$

Die Definitionen (1.3) und (1.4) modifizieren wir nun für den nichtlinearen Fall.

(1.8) Definition

Eine Folge stetiger nichtlinearer Operatoren $K_i : \overline{B}_i \to E_i$, $i \in \mathbb{N}$ heißt diskret kompakt, wenn es für jede Teilfolge $I' \subset \mathbb{N}$ und jede beschränkte Folge $\{u_i\}_{i \in I'}$ eine Teilfolge $I'' \subset I'$ und ein Element $w \in E_0$ gibt mit der Eigenschaft

$$\|K_i u_i - w\| \to 0, \quad i \in I'' \subset I' \subset \mathbb{N}.$$

(1.9) Definition

Die Folge stetiger nichtlinearer Operatoren $K_i : \overline{B}_i \to E_i$, $i \in \mathbb{N}$ heißt diskret konvergent gegen den stetigen Operator $K : \overline{B} \to E_0$, wenn für jedes $u \in \overline{B}$ und jede konvergente Folge $u_i \in \overline{B}_i$, $i \in \mathbb{N}$ gilt:

$$u_i \to u \Rightarrow K_i u_i \to Ku .$$

§ 2. Ein Konvergenzsatz für die diskrete Approximation eines Operators durch eine diskret kompakte Folge

In diesem Paragraphen stellen wir uns das folgende Problem. K habe einen Fixpunkt, $\{K_i\}_{i\in\mathbb{N}}$ sei eine diskret kompakte diskrete Approximation für K. Haben dann die Operatoren K_i für genügend großes $i\in\mathbb{N}$ einen Fixpunkt, und approximieren diese den Fixpunkt von K?

Hierzu nehmen wir an, daß K auf ∂B keinen Fixpunkt hat und die Rotation auf ∂B von Null verschieden ist. Dann gilt der folgende Konvergenzsatz.

(2.1) Hauptsatz

(1) Seien E_0, E_i, $i\in\mathbb{N}$ abgeschlossene, lineare Teilräume eines reellen Banachraumes E. $\{E_i\}_{i\in\mathbb{N}}$ sei eine diskrete Approximation von E_0.

(2) Die diskret kompakte Folge vollstetiger Operatoren $K_i : \overline{B}_i \to E_i$, $i\in\mathbb{N}$ sei diskret konvergent gegen den vollstetigen Operator $K : \overline{B} \to E_0$.

(3) K habe keinen Fixpunkt auf ∂B und genau einen Fixpunkt u_0 in B, dessen Index von Null verschieden ist.

Dann gibt es ein $i_0\in\mathbb{N}$, so daß K_i für $i \geq i_0$ einen Fixpunkt $u_i \in B_i$ hat, und jede Folge von Fixpunkten u_i konvergiert gegen den Fixpunkt u_0 von K.

Beweis:

Wir zeigen $\quad \mathrm{rot}(I-K,\partial B) = \mathrm{rot}(I-K_i,\partial B_i)$.
Da K nach Voraussetzung kompakt ist, können wir K in der üblichen Weise mit dem Leray-Schauderschen Projektionsoperator approximieren. Sei $u_0 \in B$ der Fixpunkt von K. Dann gibt es ein $\varepsilon > 0$ mit der Eigenschaft

$$(2.1.1) \qquad \bigwedge_{x\in\partial B} \|x-Kx\| \geq \varepsilon > 0 \ .$$

Andernfalls gibt es eine Folge $\{x_i\} \subset \partial B$ mit $\|x_i - Kx_i\| \to 0$. Da K kompakt und ∂B beschränkt, gibt es eine Teilfolge $\{x_i'\} \subset \{x_i\}$ mit $Kx_i' \to y$, also $x_i' \to y$. Da ∂B abgeschlossen, ist $y \in \partial B$. Die Stetigkeit von K liefert $Ky = y$, im Widerspruch zur Voraussetzung (3).

Da K kompakt ist, gibt es ein $\varepsilon/8$-Netz für $K\bar{B}$

$$K\bar{B} \subset \bigcup_{k=1}^{n_\varepsilon} K(x_k, \varepsilon/8), \quad K(x_k, \varepsilon/8) := \{x \in E_o : \|x - x_k\| < \varepsilon/8\},$$

$$x_k \in E_o .$$

Da $\{E_i\}_{i \in \mathbb{N}}$ eine diskrete Approximation für E_o ist, gibt es ein $i_o \in \mathbb{N}$, so daß für jedes k $(1 \le k \le n_\varepsilon)$ und jedes $i \ge i_o$ ein $x_k(i) \in E_i$ existiert mit

$$\|x_k(i) - x_k\| \le \frac{\varepsilon}{8} .$$

Nun definieren wir für jedes $i \ge i_o$ einen Projektionsoperator (vgl. KRASNOSEL'SKII [1964], S. 106)

$$(2.1.2) \qquad P_i z := \frac{\sum_{k=1}^{n_\varepsilon} m_k^{(i)}(z) x_k(i)}{\sum_{k=1}^{n_\varepsilon} m_k^{(i)}(z)}, \quad i=0,\ i \ge i_o,\ z \in K\bar{B},$$

wobei

$$m_k^{(i)}(z) := \begin{cases} \frac{\varepsilon}{4} - \|z - x_k(i)\|, & \text{falls } \|z - x_k(i)\| \le \frac{\varepsilon}{4} \\ 0 & \text{sonst} \end{cases}$$

$$(k=1, \ldots, n_\varepsilon) .$$

Für $x \in \bar{B}$ gilt nach Konstruktion (2.1.2)

$$(2.1.3) \qquad \|P_i Kx - Kx\| \le \frac{\varepsilon}{4}$$

und

$$P_i Kx \in E_i , \qquad i=0,\ i \ge i_o .$$

Also ist auf dem Rand von B wegen (2.1.1) und (2.1.3):

$$x \in \partial B \Rightarrow \|x-P_iKx\| \geq \|x-Kx\| - \|P_iKx-Kx\| \geq \frac{3}{4}\varepsilon .$$

Der Operator P_iK hat also auf ∂B keinen Fixpunkt, und es gilt

(2.1.4) $\quad \mathrm{rot}(I-K,\partial B) = \mathrm{rot}(I-P_iK,\partial B)$ (KRASNOSEL'SKII [1964]).

Da wir die Projektionsoperatoren geeignet gewählt haben, hat wegen $\partial B_i \subset \partial B \cap E_i$ der Operator P_iK auf ∂B_i keinen Fixpunkt. Trivial ist jeder Fixpunkt von P_iK in B_i auch Fixpunkt in B. Ist x aber ein Fixpunkt von P_iK in B, so gilt $x = P_iKx$, also $x \in E_i$ und damit $x \in B_i$. Die Fixpunkte von P_iK in B und B_i stimmen daher überein, folglich

(2.1.5) $\qquad \mathrm{rot}(I-P_iK,\partial B) = \mathrm{rot}(I-P_iK,\partial B_i)$,

also insgesamt aus (2.1.4) und (2.1.5)

(2.1.6) $\qquad \mathrm{rot}(I-K,\partial B) = \mathrm{rot}(I-P_iK,\partial B_i)$.

Nun zeigen wir noch die Homotopie der Vektorfelder $I - P_iK$ und $I - K_i$ auf ∂B_i.

Hierzu definieren wir die Homotopiefunktion

$$X_i(x,t) := t(I-K_i)x + (1-t)(I-P_iK)x, \quad i \geq i_o .$$

Es ist $X_i(x,0) = (I-P_iK)x$ und $X_i(x,1) = (I-K_i)x$, daher nach einfacher Umformung

$$X_i(x,t) = x - tK_ix - (1-t)P_iKx .$$

Wir zeigen $X_i(x,t) \neq 0$, $i \geq i_o$, $0 \leq t \leq 1$, $x \in \partial B_i$ durch Widerspruch. Es gebe also Folgen $\{t_i\} \subset [0,1]$ und $\{x_i\}$,

$x_i \in \partial B_i$ mit

$$x_i = t_i K_i x_i + (1-t_i) P_i K x_i \,. \tag{2.1.7}$$

Beachtet man die Definition (2.1.2) von P_o und P_i $(i \geq i_o)$ sowie die Stetigkeit der Norm, so folgt $\lim\limits_{i \to \infty} \sup\limits_{x \in B} \| P_i K x - P_o K x \| = 0$.

Da $\{K_i\}$ diskret kompakt, K kompakt und $\{x_i\}$ beschränkt ist, gibt es Teilfolgen $\{x_i'\} \subset \{x_i\}$ und $\{t_i'\} \subset \{t_i\}$ mit

$$t_i' \to t$$

und $\quad K_i x_i' \to w \,, \quad K x_i' \to z \,, \quad w, z \in \overline{KB} \,.$

Also konvergiert wegen (2.1.7) $\{x_i'\}$ selbst, und es ist

$$x_i' \to tw + (1-t) P_o z =: v, \quad v \in \partial B \,.$$

Da $\{K_i\}$ diskret K approximiert, folgt $K_i x_i' \to Kv$. Wegen der Stetigkeit von K ist $z = Kv$, also insgesamt

$$v = tKv + (1-t) P_o Kv.$$

Hieraus folgt mit (2.1.3) $\| v - P_o Kv \| \leq t \| Kv - P_o Kv \| \leq \frac{\epsilon}{4}$. Wegen $\| v - Kv \| \leq \| v - P_o Kv \| + \| P_o Kv - Kv \| \leq \frac{\epsilon}{2}$, (2.1.3) und $v \in \partial B$ erhalten wir einen Widerspruch zu Voraussetzung (3).

Also ist $X_i(x,t) \neq 0$ für $i \geq i_o$, $0 \leq t \leq 1$ und $x \in \partial B_i$, daher $X_i(x,t)$ eine Homotopie.

Homotope Vektorfelder haben dieselbe Rotation, also gilt $\operatorname{rot}(I - P_i K, \partial B_i) = \operatorname{rot}(I - K_i, \partial B_i)$. Zusammen mit (2.1.6) folgt daraus

$$\operatorname{rot}(I-K, \partial B) = \operatorname{rot}(I - K_i, \partial B_i) \,.$$

Also hat $I - K_i$ für $i \geq i_o$ einen Fixpunkt u_i in B_i.

Die Folge $\{u_i\} \subset B$ ist beschränkt. Da $\{K_i\}$ diskret kompakt,

gibt es eine Teilfolge $\{u_i'\} \subset \{u_i\}$ mit

$$u_i' = K_i u_i' \to u' .$$

$\{K_i\}$ konvergiert aber diskret gegen K, folglich

$$K_i u_i' \to Ku'$$

und daraus $Ku' = u'$. Da K nach Voraussetzung genau einen Fixpunkt in $\overline{B}$ hat, folgt die letzte Behauptung des Satzes.

Für das weitere wollen wir stets voraussetzen, daß $\{E_i\}_{i \in \mathbb{N}}$ eine diskrete Approximation von E_o ist.

§ 3. Konvergenz von Eigenwerten, Eigenvektoren und stetigen Zweigen

Im Anschluß an den Hauptsatz untersuchen wir den Zusammenhang zwischen Eigenwerten, Eigenvektoren und stetigen Zweigen der approximierenden Operatoren und des approximierten Operators. Aus KRASNOSEL'SKII [1964], S. 186, entnehmen wir den folgenden Satz.

(3.1) SATZ

Sei ∂B der Rand eines beschränkten Gebietes B eines Banachraumes E_o. Der Operator $K : \overline{B} \to E_o$ sei vollstetig. Ist entweder

$$0 \in B \quad \text{und} \quad \operatorname{rot}(I-K, \partial B) \neq 1$$

oder

$$0 \notin \overline{B} \quad \text{und} \quad \operatorname{rot}(I-K, \partial B) \neq 0 ,$$

dann hat K auf ∂B mindestens einen Eigenvektor u mit einem positiven Eigenwert.

Hiermit erhalten wir aus Hauptsatz (2.1) einen Konvergenzsatz für die Eigenvektoren der approximierenden Operatoren.

(3.2) SATZ

Für den Operator K seien die Voraussetzungen von Satz (3.1) erfüllt, außerdem sei u eindeutig.

Die diskret kompakte Folge vollstetiger Operatoren $K_i : \overline{B}_i \to E_i$, $i \in \mathbb{N}$ sei diskret konvergent gegen den vollstetigen Operator $K : \overline{B} \to E_o$. Dann hat K_i für genügend großes $i \geq i_o \in \mathbb{N}$ einen Eigenvektor $u_i \in \partial B_i$ mit einem positiven Eigenwert λ_i; ferner gilt für jede Folge $\{u_i\}$ von Eigenvektoren und zugehörigen Eigenwerten λ_i

$$\lim_{i\to\infty} \|u-u_i\| = 0 , \qquad \lim_{i\to\infty} |\lambda-\lambda_i| = 0 ,$$

wobei u der aufgrund der gemachten Voraussetzungen eindeutig bestimmte Eigenvektor von K mit zugehörigem Eigenwert λ ist.

Beweis:

Sei zunächst $0 \in B$ und $\mathrm{rot}(I-K,\partial B) \neq 1$.

Für genügend großes i gilt nach Satz (2.1)

$$\mathrm{rot}(I-K_i,\partial B_i) = \mathrm{rot}(I-K,\partial B) \neq 1 .$$

Mit $0 \in B$ ist auch $0 \in B_i = B \cap E_i$.

Also ist Satz (3.1) auf K_i anwendbar und liefert einen Eigenvektor $u_i \in \partial B_i$ mit positivem Eigenwert.

Ist nun $0 \notin \overline{B}$, so ist auch $0 \notin \overline{B}_i$.

Denn wegen $\partial B_i \subset \partial B \cap E_i$ hat man $0 \in \overline{B}_i \Rightarrow 0 \in B_i \cup \partial B_i \Rightarrow 0 \in B$ oder $0 \in \partial B$, also einen Widerspruch. Wiederum ist Satz (3.1) anwendbar. Die Konvergenz beweist man ähnlich wie im Hauptsatz.

Nun benötigen wir den Begriff des stetigen Zweiges (KRASNOSEL'SKII [1961]).

(3.3) Definition

Eine Menge $N \subset E$ heißt stetiger Zweig in $R(a,b,u_o) := \{u \in E : a < \|u-u_o\| < b\}$, wenn der Rand jedes Gebietes G, welches $\{u:\|u-u_o\| \leq a\}$ enthält und mit seinem Rand in $\{u:\|u-u_o\| < b\}$ enthalten ist, mit N einen nichtleeren Durchschnitt bildet.

(3.4) Definition

Ein stetiger Zweig in $R(0,b,u_o)$ heißt stetiger Zweig durch u_o.

Schreibweise:

Gilt für G und $R(a,b,u_o)$ der Zusammenhang $G \supset \{u:\|u-u_o\| \leq a\}$ und $\overline{G} \subset \{u:\|u-u_o\| < b\}$, so schreiben wir dafür abkürzend

$$G \prec R(a,b,u_o) .$$

Unter geeigneten Voraussetzungen bilden die Eigenvektoren eines nichtlinearen Operators einen stetigen Zweig durch einen isolierten Eigenwert. Hierzu zitieren wir einen Satz von KRASNOSEL'SKII [1964], S. 187.

(3.5.) SATZ

Der vollstetige Operator K habe einen Eigenwert $\lambda_o > 0$ mit zugehörigem isolierten Eigenvektor $u_o \in B$, der einen von Null verschiedenen Index hat und in $\overline{B}$ eindeutig ist. Ferner sei $0 \notin \overline{B}$. Dann gibt es ein Gebiet $R(0,b,u_o)$, in dem die Eigenvektoren einen stetigen Zweig durch u_o bilden. In diesem Fall ist das Spektrum von K stetig.

Es ist nun zu klären, in welchem Verhältnis die stetigen Zweige von approximiertem und approximierendem Operator stehen. Hierzu benötigen wir eine Folgerung aus dem Beweis des Hauptsatzes.

(3.6) Folgerung

Unter den Voraussetzungen des Hauptsatzes (2.1) gilt

$$\mathrm{rot}(I-K,\partial B^{\varepsilon}) = \mathrm{rot}(I-K_i,\partial B_i^{\varepsilon}) ,$$

wo $B^{\varepsilon}(\varepsilon>0)$ ein solches Teilgebiet von B ist, daß $\partial B^{\varepsilon} \cap \{u:\|u-u_0\|<\varepsilon\} = \emptyset$.
Der Beweis dieser Tatsache folgt wie im Beweis des Hauptsatzes, wenn man dort überall ∂B durch ∂B^{ε} bzw. ∂B_i durch $\partial B_i^{\varepsilon}$ ersetzt.

Nun können wir zeigen, daß zu jedem $a > 0$ für genügend großes i die Eigenvektoren von K_i stetige Zweige in $R(a,b,u_0) \cap E_i$ bilden und daß in einem gewissen Sinne Konvergenz der stetigen Zweige besteht.

(3.7) Lemma

Für den Operator $K : \overline{B} \to E_0$ seien die Voraussetzungen von Satz (3.5) erfüllt. Die diskret kompakte Folge vollstetiger Operatoren $K_i : \overline{B}_i \to E_i$ sei diskret konvergent gegen K.

Ist G ein Gebiet, so daß $G \prec R(a,b,u_0)$ gilt, dann gibt es ein $i(a) \in \mathbb{N}$, so daß K_i auf ∂G_i (Rand von $G_i := G \cap E_i$ in E_i) einen Eigenvektor enthält.

Beweis:

Sei $a > 0$ fest, dann ist $\partial G \cap \{u:\|u-u_0\|<a\} = \emptyset$.
Also ist aufgrund der Folgerung (3.6)

$$0 \neq \mathrm{rot}(I-K,\partial G) = \mathrm{rot}(I-K_i,\partial G_i) .$$

Da $0 \notin \overline{B} \supset \overline{G}$, ist $0 \notin \overline{G}_i$; nach Satz (3.1) folgt also die Behauptung.

Im folgenden Satz sehen wir, daß die Punkte der stetigen Zweige der approximierenden Gleichungen stets gegen Punkte der stetigen Zweige der approximierten Gleichung konvergieren.

(3.8) SATZ

Unter den Voraussetzungen von Lemma (3.7) gibt es zu jedem $\varepsilon > 0$ ein $i(\varepsilon) \in \mathbb{N}$, so daß für alle $i \geq i(\varepsilon)$ gilt:

$$\bigwedge_{G \prec R(a,b,u_0)} \bigwedge_{\substack{u_i \in \partial G_i \\ \lambda_i u_i = K_i u_i}} \inf_u \{\|u_i - u\| : \lambda u = Ku \text{ und } u \in \cup\{\partial G, G \prec R(a,b,u_0)\}\} < \varepsilon ,$$

falls $i \geq i(\varepsilon) \geq i(\varepsilon)$.

Beweis: Durch Widerspruch.

Annahme: Es gibt eine Folge $\{u_i\}$, $\lambda_i u_i = K_i u_i$, $u_i \in \partial G_i$ mit

$$\inf_u \{\|u_i - u\| : \lambda u = Ku \text{ und } u \in \cup\{\partial G, G \prec R(a,b,u_0)\}\} \geq \varepsilon .$$

Da $\{u_i\}$ beschränkt und $\{K_i\}$ diskret kompakt, gibt es eine Teilfolge $\{u_i'\} \subset \{u_i\}$ und ein w mit $K_i u_i' \to w$, also $\lambda_i u_i' \to \lambda u := w$.

Aus $u_i' \to u$ folgt wegen der diskreten Konvergenz

$$Ku_i' \to Ku, \text{ also } Ku = \lambda u .$$

Da $u_i \in \partial G_i$, ist $u \in \cup\partial G$, $G \prec R(a,b,u_0)$. Daher

$$u \in \{u : \lambda u = Ku \text{ und } u \in \cup\{\partial G, G \prec R(a,b,u_0)\}\}$$

und $u_i' \to u$, das ist aber ein Widerspruch zur Annahme.

Die Untersuchungen des ersten Kapitels sind im wesentlichen Existenz- und Konvergenzaussagen unter ziemlich allgemeinen Voraussetzungen. Die Beweise sind nicht konstruktiv und die Aussagen qualitativer Natur. Als Beispiel wollen wir den Hauptsatz auf eine nichtlineare Integralgleichung anwenden. Das folgende Ergebnis erhält auch VAINIKKO [1969] als Ergebnis einer ähnlichen Theorie, die nicht auf der Begriffsbildung von STUMMEL [1970] beruht; Beweise werden dort nicht gegeben.

(3.9) SATZ

Sei

$$\int_0^1 f(s)ds = \sum_{k=0}^{n} A_k^{(n)} f(s_k) + R_n(f) ,$$

$$\bigwedge_{f\in C[0,1]} \lim_{n\to\infty} R_n(f) = 0$$

eine für stetige Funktionen konvergente Quadraturformel, die für Polynome vom Höchstgrad n exakt ist. Kollokation mit dieser Quadraturformel ergibt, angewandt auf eine Urysonsche Integralgleichung

$$(3.9.1) \qquad x(s) = \int_0^1 K(s,t,x(t))dt, \quad 0\le s\le 1$$

$$K \in C([0,1]^2 \times \mathbb{R}) ,$$

ein nichtlineares Gleichungssystem

$$(3.9.2) \qquad a_i^{(n)} = \sum_{k=0}^{n} A_k^{(n)} K(s_i^{(n)}, s_k^{(n)}, a_k^{(n)}) , \quad i=0,\dots,n$$

für den Vektor $\underline{a}^{(n)} := (a_1^{(n)},\dots,a_n^{(n)})'$.

Hat Gleichung (3.9.1) in einem offenen beschränkten Gebiet B genau einen Fixpunkt x_o, dessen Index von Null verschieden ist, und keinen Fixpunkt auf ∂B, dann hat das System (3.9.2) für genügend großes n eine Lösung $\underline{a}^{(n)}$ und es gilt

$$\lim_{n\to\infty} \max_{o\le i\le n} |a_i^{(n)} - x_o(s_i)| = 0 .$$

Beweis:

Sei $E_o := E := C[0,1]$, $E_n :=$ {Polynome mit Höchstgrad n}; die Räume seien mit der T-Norm versehen. Dann ist $\{E_n\}$ eine diskrete Approximation von E_o.

$$Kx(s) := \int_o^1 K(s,t,x(t))dt,$$

$$K_n x(s) := \sum_{k=o}^{n} A_k^{(n)} K(s,s_k,x(s_k)) .$$

Nach Voraussetzung ist $R_n f = 0$ für $f \in E_n$.

Sei $\{x_n\}$ mit $x_n \in E_n$ beschränkt, dann ist $K_n x_n = Kx_n$, und da K vollstetig, $\{K_n x_n\}$ kompakt; also ist $\{K_n\}$ diskret kompakt. Sei $x_n \to x$, dann ist $K_n x_n = Kx_n \to Kx$ wegen der Stetigkeit von K, also ist $K_n \to K$ diskret konvergent.

Der Hauptsatz ist anwendbar, also sind die Gleichungen

$$K_n x_n = x_n \tag{3.9.3}$$

für genügend großes n lösbar. Gleichung (3.9.3) bedeutet aber

$$\sum_{k=o}^{n} A_k^{(n)} K(s,s_k,x_n(s_k)) = x_n(s) , \quad o\le s\le 1 . \tag{3.9.4}$$

Da ein Polynom aus E_n durch seine Werte an den $n+1$ verschiedenen Stellen $s_o < s_1 < \ldots < s_n$ eindeutig bestimmt ist, ist

(3.9.4) äquivalent mit dem nichtlinearen Gleichungssystem

$$(3.9.5) \qquad \sum_{k=0}^{n} A_k^{(n)} K(s_i, s_k, a_k^{(n)}) = a_i^{(n)} \quad (i=0,\ldots,n) .$$

Die Lösung $\underline{a}^{(n)}$ von (3.9.5) bestimmt eindeutig eine Lösung x_n von (3.9.4), für welche gilt

$$0 = \lim_{n\to\infty} \|x_n - x_o\| = \lim_{n\to\infty} \max_{o\le s\le 1} |x_n(s) - x_o(s)| \ge$$

$$\ge \lim_{n\to\infty} \max_{o\le i\le n} |a_i^{(n)} - x_o(s_i)| ,$$

also die Behauptung.

Um zu quantitativen Aussagen, insbesondere zu Fehlerabschätzungen zu gelangen, muß man weiter auf die Struktur der zu betrachtenden Operatoren eingehen. Wir wollen hier Differenzierbarkeitseigenschaften fordern, nicht zuletzt auch um bei der praktischen Durchführung von Beispielen genügend rasch konvergente Verfahren, wie z.B. das Newtonverfahren, anwenden zu können. Das nächste Kapitel behandelt eine derartige Fragestellung.

ZWEITES KAPITEL

KONVERGENZ DIFFERENZIERBARER OPERATOREN UND NEWTON-VERFAHREN

§ 4. Ein Konvergenzsatz für die Approximation von Gleichungen der Form $Kx = 0$

In diesem Kapitel wollen wir Folgen differenzierbarer Operatoren betrachten. Es werden also keine Kompaktheitsforderungen gestellt. An die Stelle der diskreten Konvergenz tritt eine schwächere Konvergenzforderung. Zusätzlich benötigen wir aber die Konvergenz von Ableitungen. Aussagen ähnlicher Art unter der zusätzlichen Voraussetzung der "kollektiven Kompaktheit" (ein Spezialfall der diskreten Kompaktheit im ersten Kapitel) der Folge der ersten Fréchetableitung der Näherungsoperatoren erhält MOORE [1968].

Als Ausgangspunkt formulieren wir einen Existenzsatz von KANTOROWITSCH (RALL [1969], S. 135).

Wir legen wieder einen Banachraum E zugrunde und definieren

$$S(u,r) := \{x \in E : \|u-x\| < r\} , \quad u \in E, \; r>0 .$$

(4.1) SATZ

Der Operator $K : D \to E$ sei in $S(x_0,R) \subset E$ zweimal differenzierbar (im Sinne von Fréchet). Die Inverse des Operators $G := K'(x_0)$ sei beschränkt und es gelte

(4.1.1) $$\|G^{-1}\| \leq B_0 ;$$

(4.1.2) $$\|x_1-x_0\| \leq \eta_0 , \text{ wobei } x_1 := x_0 - G^{-1}Kx_0 ;$$

(4.1.3) $\bigwedge_{x\in\bar{S}(x_0,r)} \|K''(x)\| \le M , \quad r < R ;$

(4.1.4) $0 < h := B_0 r_0 M < \frac{1}{2} ;$

(4.1.5) $r^- < r < r^+, \quad \text{wobei} \quad r^{\pm} := \frac{1 \pm \sqrt{1-2h}}{h} r_0 .$

Dann konvergiert das Newton-Verfahren

$$x_{n+1} := x_n - [K'(x_n)]^{-1} Kx_n, \quad n=1,2,\ldots$$

gegen die in $\bar{S}(x_0,r)$ eindeutig bestimmte Lösung x^* der Gleichung

$$Kx = 0 .$$

Es gilt die Abschätzung

$$\|x^* - x_0\| \le r^- .$$

Wir wollen an dieser Stelle auch gleich eine Verallgemeinerung dieses Satzes angeben, die wir später benötigen werden (KANTOROWITSCH [1964]).

(4.2) Korollar

Der Operator $K : E \to E$ sei in $S(x_0,R) \subset E$ zweimal differenzierbar. Es gebe einen linearen stetigen Operator $G : E \to E$ mit den folgenden Eigenschaften:

(4.2.1) $\|GKx_0\| \le r_0 ;$

(4.2.2a) $\|GK'(x_0) - I\| \le d ;$

(4.2.2.b) $\|K'(x_0)G - I\| \le d ;$

(4.2.3) $$\bigwedge_{x\in\overline{S}(x_0,r)} \|GK''(x)\| \le M \,, \quad r < R \,;$$

(4.2.4) $$h := \frac{M\eta_0}{(1-d)^2} < \frac{1}{2} \,;$$

(4.2.5) $$r^- < r < r^+, \quad \text{wobei} \quad r^{\pm} := \frac{1\pm\sqrt{1-2h}}{h}\,\eta_0 \,.$$

Dann gilt die Aussage von Satz (4.1).

(4.3) Bemerkung: Bedingung (4.2.2b) wird bei KANTOROWITSCH nicht gefordert. MORRE [1968] gibt jedoch ein Beispiel an, das zeigt, daß (4.2.1), (4.2.2a), (4.2.3), (4.2.4) und (4.2.5) nicht hinreichend sind. Fordert man jedoch z.B. zusätzlich (4.2.2b), läßt sich der Satz behaupten.

Nun betrachten wir eine Folge von Operatoren $K_i : E \to E$, die ebenfalls in $S(x_0,R)$ zweimal differenzierbar sind und den Operator K in einem geeigneten Sinne approximieren. Die Approximationsforderungen stellen uns sicher, daß sich die Voraussetzungen über K für genügend großes $i \in \mathbb{N}$ auf K_i übertragen, und dadurch Existenz einer Lösung der Gleichung $K_i x_i = 0$ sowie die Konvergenz des Newton-Verfahrens erreicht wird.

(4.4) Hauptsatz

$K : E \to E$ erfülle die Voraussetzungen von Satz (4.1). Die Operatoren $K_i : E \to E$ seien in $S(x_0,R)$ zweimal differenzierbar und genügen den folgenden Approximationsbedingungen:

(4.4.1) $$\bigwedge_{x\in\overline{S}(x_0,r)} \lim_{i\to\infty} \|K_i x - Kx\| = 0 \,;$$

(4.4.2) $$\lim_{i\to\infty} \|K_i'(x_0) - K'(x_0)\| = 0 \,;$$

(4.4.3) $$\bigwedge_{\varepsilon>0} \bigvee_{i(\varepsilon)\in\mathbb{N}} \bigwedge_{i\geq i(\varepsilon)} \bigwedge_{x\in\overline{S}(x_0,r)} \|K_i'(x)-K'(x)\| < \varepsilon .$$

Dann gibt es ein $i_0 \in \mathbb{N}$, so daß die Gleichung

$$K_i x_i = 0$$

für alle $i \geq i_0$ eine in $\overline{S}(x_0,r)$ eindeutig bestimmte Lösung x_i^* hat. Die Folge der Lösungen $\{x_i^*\}_{i\geq i_0}$ konvergiert gegegen die in $\overline{S}(x_0,r)$ eindeutig bestimmte Lösung x^* der Gleichung $Kx = 0$.

Mit der Anfangsnäherung x_0 konvergiert auch das Newton-Verfahren

$$x_{n+1} = x_n - [K_i'(x_n)]^{-1} K_i x_n , \quad n=0,1,2,...$$

für jedes $i \geq i_0$.

Beweis:

Wegen (4.4.2) gibt es ein $i_1 \in \mathbb{N}$

$$\|G_i-G\| \leq \|G^{-1}\|^{-1} \quad (i\geq i_1),$$

wobei $G := K'(x_0)$, $G_i := K_i'(x_0)$. Also existiert G_i^{-1} mit

$$\|G_i^{-1}-G^{-1}\| \leq \frac{\|G^{-1}\|^2\|G_i-G\|}{1-\|G^{-1}\|\|G_i-G\|} ;$$

daher gilt

(4.4.4) $$\lim_{i\to\infty} \|G_i^{-1}-G^{-1}\| = 0 .$$

Sei $\varepsilon > 0$ beliebig, dann ist wegen (4.4.4) und (4.1.1) für $i \geq i_2(\varepsilon) \geq i_1$

$$\|G_i^{-1}\| \leq B_0 + \varepsilon .$$

Weiter gilt die Abschätzung

$$\|G_i^{-1}K_i x_0\| \leq \|G_i^{-1}K_i x_0 - G_i^{-1}Kx_0\| + \|G_i^{-1}Kx_0 - G^{-1}Kx_0\| + \|G^{-1}Kx_0\|$$

$$\leq (B_0+\varepsilon)\|K_i x_0 - Kx_0\| + \|G_i^{-1}-G^{-1}\|\|Kx_0\| + \eta_0 ,$$

also unter Verwendung von (4.4.1) und (4.4.4) für $i \geq i_3(\varepsilon) \geq i_2(\varepsilon)$

$$\|G_i^{-1}K_i x_0\| \leq \eta_0 + \varepsilon .$$

Eine Abschätzung von $K_i''(x)$ erhalten wir aus

$$\|K_i''(x)\| \leq \|K_i''(x) - K''(x)\| + \|K''(x)\| \leq M + \varepsilon, \quad x \in \bar{S}(x_0, r) ,$$

falls $i \geq i_4(\varepsilon) \geq i_3$ und die Voraussetzungen (4.1.3) und (4.4.3) beachtet werden.

Wir definieren $h(\varepsilon) := (B_0+\varepsilon)(\eta_0+\varepsilon)(M+\varepsilon)$. $h(\varepsilon)$ ist monoton wachsend mit $h(0) = h < \frac{1}{2}$ und stetig. Sei $h(\varepsilon_2) = \frac{1}{2}$, dann betrachten wir in $[0, \varepsilon_2]$ die Funktion

$$r^-(\varepsilon) := (1-\sqrt{1-2h(\varepsilon)})(\eta_0+\varepsilon)/h(\varepsilon) .$$

Diese Funktion ist in $[0, \varepsilon_2)$ streng monoton wachsend.

Wegen $r^-(0) = r^- < r$ und der Stetigkeit von $r^-(\varepsilon)$ gibt es ein $\varepsilon_1 \in (0, \varepsilon_2)$ mit $r^-(\varepsilon) < r$ für $\varepsilon \in [0, \varepsilon_1]$.

Die Funktion

$$r^+(\varepsilon) := (1+\sqrt{1-2h(\varepsilon)})(\eta_0+\varepsilon)/h(\varepsilon)$$

ist in $[0,\varepsilon_1]$ stetig mit $r^+(0) = r^+ > r$. Also gibt es ein $\varepsilon_0 \in (0,\varepsilon_1]$ mit $r_1(\varepsilon) > r$ für $\varepsilon \in [0,\varepsilon_0]$.

Sei nun $\varepsilon \in [0,\varepsilon_0]$ und $i_0 := i_4(\varepsilon)$, dann sind für den Operator K_i mit $i \geq i_0$ die Voraussetzungen des Satzes (4.1) erfüllt:

$$\|G_i^{-1}\| \leq B_0 + \varepsilon$$

$$\|x_1 - x_0\| \leq \eta_0 + \varepsilon, \text{ wobei } x_1 := x_0 - G_i^{-1}K_i x_0$$

$$\bigwedge_{x\in\bar{S}(x_0,r)} \|K_i''(x)\| \leq M + \varepsilon\ , \quad r < R$$

$$0 < h(\varepsilon) = (B_0+\varepsilon)(\eta_0+\varepsilon)(M+\varepsilon) < \frac{1}{2}$$

$$r^-(\varepsilon) < r < r^+(\varepsilon)\ .$$

Also ist für genügend große i Satz (4.1) auf K_i anwendbar und liefert eine eindeutig bestimmte Lösung x_i^* in $\bar{S}(x_0,r)$.

Es bleibt noch die Konvergenz $\|x_i^* - x^*\| \to 0$ zu zeigen. Hierzu definieren wir $\bar{K} := I - G^{-1}K$, $\bar{K}_i := I - G_i^{-1}K_i$. Dann gilt $\bar{K}(\bar{S}(x_0,r^-)) \subset \bar{S}(x_0,r^-)$, denn sei $\bar{x} \in \bar{S}$, so ist $\|\bar{x}-x_0\| \leq r^-$, also:

$$\begin{aligned}
\|\bar{K}\bar{x} - x_0\| &= \|\bar{x} - G^{-1}K\bar{x} - x_0\| = \|\bar{x} - x_0 + G^{-1}Kx_0 - G^{-1}K\bar{x} - G^{-1}Kx_0\| \\
&\leq \|G^{-1}\{K'(x_0)(\bar{x}-x_0) + Kx_0 - K\bar{x}\}\| + \|G^{-1}Kx_0\| \\
&\leq \frac{1}{2} \sup_{0<h<1} \|G^{-1}K''[x_0 + h(\bar{x}-x_0)]\| \|x-\bar{x}\|^2 + \eta_0 \\
&\leq \frac{1}{2} MB_0 (r^-)^2 + \eta_0 = \eta_0 + \frac{1}{2}\frac{h}{\eta_0}\frac{(1-\sqrt{1-2h})^2}{h^2}\eta_0^2 = r^-\ .
\end{aligned}$$

Analog zeigt man $\bar{K}_i(\bar{S}(x_0,r^-(\varepsilon))) \subset \bar{S}(x_0,r^-(\varepsilon))$.

$\bar{K}_i$ ist kontrahierend auf $\bar{S}(x_0,r^-(\varepsilon))$, denn seien $x_1,x_2 \in \bar{S}(x_0,r^-(\varepsilon))$, so folgt

$$\|\bar{K}_i x_1 - \bar{K}_i x_2\| \leq \sup_{0<h<1} \|\bar{K}_i'[x_1+h(x_2-x_1)]\| \|x_1-x_2\| .$$

Für $x \in \bar{S}(x_0,r^-(\varepsilon))$ gilt jedoch

$$\bar{K}_i'(x) = I - G_i^{-1}K_i'(x) = G_i^{-1}[K_i'(x_0)-K_i'(x)], \text{ also}$$

$$\|\bar{K}_i'(x)\| \leq (B_0+\varepsilon)(M+\varepsilon)r^-(\varepsilon) = 1 - \sqrt{1-2h(\varepsilon)} < 1 .$$

Abkürzend schreiben wir für die Kontraktionskonstante $c_\varepsilon := 1 - \sqrt{1-2h(\varepsilon)}$. Analog zeigt man, daß $\bar{K}$ kontrahierend ist mit der Kontraktionskonstanten $c := 1 - \sqrt{1-2h}$.

Nach dem Banachschen Fixpunktsatz haben $\bar{K}$ und $\bar{K}_i$ genau einen Fixpunkt $\bar{x}$ bzw. $\bar{x}_i$ in $\bar{S}(x_0,r^-)$ bzw. $\bar{S}(x_0,r^-(\varepsilon))$. Unter den gemachten Voraussetzungen gilt also wegen der Eindeutigkeit $\bar{x} = x^*$ und $\bar{x}_i = x_i^*$.

Nun schätzen wir ab

$$\|x^*-x_i^*\| = \|\bar{K}x^*-\bar{K}_i x^*\| + \|\bar{K}_i x^*-\bar{K}_i x_i^*\| \leq \|G^{-1}Kx^*-G_i^{-1}K_i x^*\| +$$

$$+ c_\varepsilon\|x^*-x_i^*\| \leq \|G^{-1}\|\|Kx^*-K_i x^*\| + \|(G^{-1}-G_i^{-1})K_i x^*\| +$$

$$+ c_\varepsilon\|x^*-x_i^*\| ; \text{ wegen } c_\varepsilon < 1 \text{ folgt}$$

$$(4.4.5) \qquad \|x^*-x_i^*\| \leq \frac{1}{1-c_\varepsilon}\{\|G^{-1}\|+\|G^{-1}-G_i^{-1}\|\}\|Kx^*-K_i x^*\| .$$

Die Gleichungen (4.4.1) und (4.4.4) liefern $x^*-x_i^* \to 0$

Da die Lösung x^* im allgemeinen nicht bekannt ist, wählt man als Ausgangspunkt für Fehlerbetrachtungen besser die folgende Abschätzung:

$$\|x^*-x_i^*\| \leq \|\bar{K}x^*-\bar{K}x_i^*\| + \|\bar{K}x_i^*-\bar{K}_i x_i^*\| \leq c\|x^*-x_i^*\| + B_0\|K_i x_i^*-Kx_i^*\| ,$$

$$\text{also } \|x^*-x_i^*\| \leq \frac{B_0}{1-c}\|Kx_i^*\| .$$

Obwohl wir später Fehlerabschätzungen mit anderen Mitteln gewinnen werden, wollen wir hier auch eine quantitative Formulierung des Hauptsatzes geben. Aus Gründen der einfacheren Schreibweise wollen wir dann im folgenden die qualitative Formulierung beibehalten.

(4.5) Korollar

Der Operator $K : E \to E$ erfülle die Voraussetzungen von Satz (4.1). $\tilde{K} : E \to E$ sei in $S(x_0,R)$ zweimal differenzierbar und es gelte dort

(4.5.1) $\quad \|\tilde{K}x-Kx\| \le d_1(x)$

(4.5.2) $\quad \|\tilde{K}'(x_0)-K'(x_0)\| \le d_2 < \|G^{-1}\|^{-1}$

(4.5.3) $\quad \|\tilde{K}''(x)-K''(x)\| \le d_3 \quad$ für $\quad x \in \bar{S}(x_0,r),\ r < R$

4.5.4) $\quad \|x_1-x_0\| \le \|G^{-1}\|\|Kx_0\| \le \eta_0 .$

Mit den Fehlergrößen

$$\tilde{B}_0 := \frac{B_0}{1-B_0 d_2} \cdot \quad \tilde{\eta}_0 := \frac{\eta_0+B_0 d_1(x_0)}{1-B_0 d_2} , \quad \tilde{M} := M+d_3$$

gelte

$$\tilde{h}_0 := \tilde{B}_0 \tilde{\eta}_0 \tilde{M} < \frac{1}{2} \quad \text{und} \quad \tilde{r}^- < r < \tilde{r}^+ \quad \text{(analog definiert).}$$

Dann hat $\tilde{K}\tilde{x} = 0$ eine in $\bar{S}(x_0,r)$ eindeutig bestimmte Lösung $\tilde{x}^*$ und es gilt die Abschätzung

$$\|\tilde{x}^*-x^*\| \le \frac{B_0}{1-B_0 K r^-} \, d_1(\tilde{x}^*) .$$

Mit der Anfangsnäherung x_0 konvergiert auch das Newton-Verfahren

$$x_{n+1} := x_n - [\tilde{K}'(x_n)]^{-1} \tilde{K}x_n, \quad n=0,1,2,\ldots$$

Insgesamt sehen wir also, daß das Newton-Verfahren gegen geeignete kleine Störungen des Operators unempfindlich ist. Vertauscht man die Rolle von P und $\tilde{P}$, kann man natürlich auch von der Lösbarkeit der Näherungsgleichung auf die Existenz einer Lösung der exakten Gleichung schließen. Korollar (4.5) gilt dann entsprechend.

§ 5. Der Konvergenzsatz für Gleichungen zweiter Art:

$$x = Qx + f$$

Im Hinblick auf unser Ziel - Lösung Hammersteinscher Gleichungen durch Kollokation - wollen wir den Hauptsatz nun auf Gleichungen der Form

$$x = Qx + f \ , \quad x,f \in E \ ,$$

$Q : E \to E$ nichtlinear, anwenden.

Zur Definition der Näherungsoperatoren betrachten wir noch lineare Operatoren $H_i : E \to E_i$, die E auf abgeschlossene lineare Teilräume E_i abbilden. Als Näherungsgleichungen erklären wir in E_i:

$$x_i = H_i Q x_i + H_i f \ .$$

Der Operator $K : E \to E$ sei nun gemäß $Kx := x - Qx - f$, $x \in E$ definiert, dann gilt der folgende Satz.

(5.1) SATZ

Der Operator $Q : E \to E$ sei in $S(x_0,R)$ zweimal differenzierbar und K genüge den Voraussetzungen von Satz (4.1). Für die Folge $\{H_i\}_{i\in\mathbb{N}}$ linearer beschränkter Operatoren gelte

$$(5.1.1)\quad \begin{cases} \lim\limits_{i\to\infty} \|(I-H_i)Qx\| = 0 & \text{für } x\in\bar{S}(x_0,r) \\ \lim\limits_{i\to\infty} \|(I-H_i)x_0\| = 0 \\ \lim\limits_{i\to\infty} \|(I-H_i)f\| = 0\,, & f \text{ fest} \end{cases}$$

$$(5.1.2)\quad \lim_{i\to\infty} \|(I-H_i)Q'(x_0)\| = 0$$

$$(5.1.3)\quad \lim_{i\to\infty} \|(I-H_i)Q''(x)\| = 0 \quad \text{gleichmäßig in } \bar{S}(x_0,r)\,.$$

Dann gilt für die auf E_i gemäß $K_ix_i := x_i - H_iQx_i - H_if$ definierten Näherungsoperatoren die Aussage des Hauptsatzes (4.4), wenn man als Anfangsnäherung $\bar{x}_0 := H_ix_0$ wählt.

Beweis:

Es ist wegen (5.1.1) für $x \in \bar{S}$

$$\|K_ix-Kx\| \le \|(I-H_i)Qx\| + \|(I-H_i)f\| \to 0, \quad i\to\infty\,.$$

Da $K'(x_0) = I-Q'(x_0)$ und $K_i'(x_0) = I-H_iQ'(x_0)$, gilt mit (5.1.2)

$$\|K_i'(x_0)-K'(x_0)\| \le \|(I-H_i)Q'(x_0)\| \to 0, \quad i\to\infty\,.$$

Schließlich ist $K''(x) = -Q''(x)$ und $K_i''(x) = -H_iQ''(x)$, also wegen (5.1.3)

$$\|K''(x)-K_i''(x)\| = \|(I-H_i)Q''(x)\| \to 0, \quad i\to\infty\,,$$

gleichmäßig in $\bar{S}$.

Das Newton-Verfahren ist bezüglich der Anfangsnäherung stabil, denn es konvergiert für jede Anfangsnäherung $\bar{x}_o \in \bar{S}(x_o,r)$ mit $\|x_o-\bar{x}_o\| \leq \frac{1-2h}{4MB_o}$, also für i genügend groß auch für $\bar{x}_o = H_i x_o$. Damit ist der Satz bewiesen.

Bei der praktischen Lösung der Näherungsgleichungen ist es oft zweckmäßig, diese nicht in den Teilräumen E_i (z.B. $E_i = \mathrm{span}(1,x,x^2,\ldots,x^i)$), sondern in dazu isomorphen Teilräumen F_i (z.B. Teilräume eines $\mathbb{R}^i$) eines Banachraumes F zu lösen.

Es seien also $h_i : E_i \to F_i$ stetige lineare Operatoren mit beschränkter Inversen h_i^{-1}. Dann betrachten wir in F_i die Näherungsgleichung

$$y_i = h_i H_i Q h_i^{-1} y_i + h_i H_i f \ , \quad y_i \in F_i \ ,$$

was gleichbedeutend ist mit der Gleichung

$$L_i y_i = 0 \ ,$$

wenn $L_i : F_i \to F_i$ gemäß $L_i y_i := y_i - h_i H_i Q h_i^{-1} y_i - h_i H_i f$ definiert wird. Für die Lösbarkeit von $L_i y_i = 0$ ergibt sich aus Satz (5.1) folgende Aussage.

(5.2) SATZ

Es seien die Voraussetzungen von Satz (5.1) erfüllt und für die Anfangsnäherung x_o werden die Größen $x_{oi} := h_i H_i x_o$ und $\varepsilon_i := \|(I-H_i)x_o\|$ definiert. Falls

$$\lim_{i\to\infty} \varepsilon_i = 0 \ , \tag{5.2.1}$$

gibt es ein $i_o \in \mathbb{N}$, so daß für alle $i \geq i_o$ die Gleichung

$$L_i y_i = 0 \tag{5.2.2}$$

in der Kugel $\overline{S}(x_{oi}, \|h_i\|(r+\varepsilon_i))$ eine Lösung y_i^* hat, die in $\overline{S}(x_{oi}, \|h_i^{-1}\|^{-1}(r-\varepsilon_i))$ eindeutig ist, und es gilt die Konvergenz

$$\lim_{i\to\infty} \|h_i^{-1} y_i^* - x^*\| = 0 .$$

Mit der Anfangsnäherung $y_o := h_i H_i x_o$ konvergiert das Newton-Verfahren

$$y_{n+1} = y_n - [L_i'(y_n)]^{-1} L_i y_n , \qquad n=0,1,2,\ldots$$

für jedes $i \geq i_o$ gegen y_i^* .

Beweis:

Der Operator L_i ist in $S_i := S(x_{oi}, R\|h_i^{-1}\|^{-1})$ zweimal differenzierbar.

Denn für $y_i \in S_i$ ist $h_i^{-1} y_i \in S(x_o, R)$, da $\|x_{oi} - h_i^{-1} y_i\| \leq \|h_i^{-1}\| \|y_i - x_{oi}\| < R$. Wir wollen die Ableitungen von L_i angeben:

$L_i'(y_i)$ angewandt auf ein Element $z_i \in F_i$ ist

$$L_i'(y_i) z_i = z_i - h_i Q_i'(h_i^{-1} y_i) h_i^{-1} z_i .$$

$L_i''(y_i)$ angewandt auf ein Paar $(z_i, z_i') \in F_i \times F_i$ ergibt

$$L_i''(y_i)(z_i, z_i') = -h_i Q_i''(h_i^{-1} y_i)(h_i^{-1} z_i, h_i^{-1} z_i') .$$

Hierbei ist $Q_i := H_i Q$, also $Q_i' = H_i Q'$ und $Q_i'' = H_i Q''$.

Aufgrund von Satz (5.1) hat $K_i x_i = 0$ eine Lösung x_i^*. Trivial ist $y_i^* := h_i x_i^*$ Lösung von (5.2.2), für die gilt

$$\|y_i^* - x_{oi}\| = \|h_i(x_i^* - H_i x_o)\| \le \|h_i\|\{\|x_i^* - x_o\| + \|(I-H_i)x_o\|\} \le$$

$$\le \|h_i\|(r+\varepsilon_i) ,$$

also liegt y_i^* in der behaupteten Kugel.

Sei $y_i^{**} \in \bar{S}(x_{oi}, \|h_i^{-1}\|^{-1}(r-\varepsilon_i))$ eine weitere Lösung von (5.2.2), so gilt für $x_i^{**} := h_i^{-1} y_i^{**}$

$$\|x_i^{**} - x_o\| \le \|h_i^{-1} y_i^{**} - H_i x_o\| + \|(I-H_i)x_o\| \le \|h_i^{-1}\|^{-1}\|y_i^{**} - x_{oi}\| +$$

$$+ \varepsilon_i \le r .$$

Also ist x_i^{**} eine Lösung von $K_i x_i = 0$ in $\bar{S}(x_o, r)$, für welche wegen der eindeutigen Lösbarkeit $x_i^{**} = x_i^*$ gelten muß, also ist auch $y_i^{**} = y_i^*$.

Für den letzten Teil der Behauptung genügt es zu zeigen, daß zwischen den Newton-Iterierten $x_{n+1} = x_n - [K_i'(x_n)]^{-1} K_i x_n$ und $y_{n+1} = y_n - [L_i'(y_n)]^{-1} L_i y_n$ der Zusammenhang

$$y_n = h_i x_n \qquad n=0,1,2,\ldots \tag{5.2.3}$$

besteht.

Beachten wir, daß in Satz (5.1) statt x_o die Anfangsnäherung $H_i x_o$ gewählt wurde, so gilt (5.2.3) für $n = 0$.

Sei nun (5.2.3) bis einschließlich n richtig, so folgt

$$y_{n+1} = y_n - [L_i'(y_n)]^{-1} L_i y_n = h_i x_n - [L_i'(h_i x_n)]^{-1} L_i h_i x_n .$$

Wie man leicht nachrechnet, ist $L_i'(h_i x_n) h_i = h_i K_i'(x_n)$ und $L_i h_i x_n = h_i K_i x_n$, also $y_{n+1} = h_i x_n - h_i [K_i'(x_n)]^{-1} K_i x_n = h_i x_{n+1}$.

(5.2.3) gilt daher für alle n, womit der Satz bewiesen ist.

(5.3) Bemerkung Gibt es ein $i_o \in \mathbb{N}$, so daß $x_o \in E_i$ für $i \geq i_o$ (etwa bei aufsteigenden Teilräumen), und sind die H_i Projektionen, dann gilt Satz (5.2) für $i \geq i_o$ mit $\varepsilon_i = 0$.

§ 6. Der Konvergenzsatz für Hammersteinsche Gleichungen

$$x = LNx + f$$

Wir wollen nun den Operator Q weiter spezialisieren. Er möge die spezielle Form

$$Q = LN$$

haben, wobei L ein linearer und N ein nichtlinearer Operator sein soll. Das führt zur Lösung einer sogenannten Hammersteinschen Gleichung

$$Kx \equiv x - LNx - f = 0 .$$

Für praktische Bedürfnisse ist eine Approximationsgleichung der Form $K_i x_i \equiv x_i - H_i LNx_i - H_i f = 0$ noch unbefriedigend. Man möchte vielmehr L und N für sich diskretisieren können, das heißt eine Näherungsgleichung $x_i - H_i LH_i Nx_i - H_i f = 0$ lösen. Solche Näherungen betrachtet KOLODNER [1964] und gewinnt einen Konvergenzsatz für Hammersteinsche Gleichungen mit Hilbertraummethoden. Seine Aussagen und Methoden sind mit unseren nicht vergleichbar.

Wir wollen nun Näherungsoperatoren M_i in E_i betrachten, die folgendermaßen definiert sind:

$$M_i x_i := x_i - H_i LH_i Nx_i - H_i f, \quad x_i \in E_i .$$

Analog zu L_i definieren wir den zugehörigen Operator in F_i:

$$N_i y_i = y_i - h_i H_i LH_i N h_i^{-1} y_i - h_i H_i f .$$

Wir untersuchen die Lösbarkeit von $M_i x_i = 0$ und $N_i y_i = 0$ und die Konvergenz der Lösungen.

Im folgenden wollen wir stets voraussetzen, daß der lineare Operator $L : E \to E$ stetig ist, und daß der nichtlineare Operator $N : S(x_0,R) \to E$ in $\overline{S}(x_0,r)$ $(r<R)$ eine beschränkte zweite Ableitung hat, d.h.

$$\bigvee_{m''>0} \bigwedge_{x\in\overline{S}(x_0,r)} \|N''(x)\| \le m'' .$$

Trivialerweise ist dann auch N' durch eine Konstante m' und N durch eine Konstante m beschränkt in $\overline{S}(x_0,r)$.

Unter diesen Voraussetzungen gilt der folgende Satz.

(6.1) SATZ

Der für $x \in E$ gemäß $Kx := x - LNx - f$ definierte Operator erfülle die Voraussetzungen von Satz (4.1). Mit $D_i := I - H_i$ seien für $i \to \infty$ die folgenden Aussagen richtig:

(6.1.1) $\quad \|D_i L\| \to 0 , \quad \|D_i f\| \to 0 , \quad \|D_i x_0\| \to 0$

(6.1.2a) $\quad \|H_i L D_i\| \|D_i f\| \to 0$
$\quad\quad\quad \|H_i L D_i\| \|D_i L\| \to 0$

(6.1.2b) $\quad \|H_i L D_i\| \|D_i f\| \|H_i\|^2 \to 0$
$\quad\quad\quad \|H_i L D_i\| \|D_i L\| \|H_i\|^2 \to 0$

(6.1.3a) $\quad \|H_i L D_i N x^*\| \to 0$

(6.1.3b) $\quad \|H_i L D_i N x^*\| \|H_i\|^2 \to 0$

(6.1.4) $\quad \|H_i L D_i N'(x^*)\| \to 0 .$

Dann gibt es ein $i_0 \in \mathbb{N}$, so daß für $i \geq i_0$ die Gleichung

$$M_i x_i = x_i - H_i L H_i N x_i - H_i f$$

in E_i eine Lösung x_i^o hat. Ist x_i^* Lösung von $K_i x_i = 0$, so gilt die Konvergenz

$$\|x_i^o - x_i^*\| \to 0, \quad i \to \infty .$$

Gilt überdies

$$\|H_i\| \to \infty , \quad i \to \infty , \tag{6.1.5}$$

so ist x_i^o lokal eindeutig.

Beweis:

Wir zeigen, daß die Voraussetzungen von Satz (5.1) erfüllt sind.

Wegen $\|D_i Q x\| \leq m \|D_i L\|$, $\|D_i Q'(x_o)\| \leq m' \|D_i L\|$ und $\|D_i Q''(x)\| \leq m'' \|D_i L\|$ sowie (6.1.1) gelten (5.1.1), (5.1.2), (5.1.3) und damit Satz (5.1).

Sei nun x_i^* für $i \geq i_0$ die Lösung von $K_i x_i = 0$, dann ist die Folge der stetigen linearen Operatoren $G_i := [K_i'(x_i^*)]^{-1}$ gleichmäßig beschränkt. Denn es ist

$$\|K_i'(x_i^*) - K'(x^*)\| \leq \|K_i'(x_i^*) - K'(x_i^*)\| + \|K'(x_i^*) - K'(x^*)\|$$

$$\leq \|D_i L\| m' + m'' \|x_i^* - x^*\| , \quad \text{also}$$

$\|K_i'(x_i^*) - K'(x^*)\| \to 0$, $i \to \infty$. Für die Norm von G_i gilt daher

$$\|G_i\| \leq \frac{\|[K'(x^*)]^{-1}\|}{1 - \|[K'(x^*)]^{-1}\| \, \|K_i'(x_i^*) - K'(x^*)\|} \leq B .$$

Um die Lösbarkeit von $M_i x_i = 0$ zu zeigen, wollen wir Korollar (4.2) auf diese Gleichung anwenden. Als Anfangsnäherung wählen

wir x_i^* und als genügend gute Pseudoinverse für $M_i'(x_i^*)$ definieren wir $G_i := [K_i'(x_i^*)]^{-1}$. Nun verifizieren wir der Reihe nach die Voraussetzungen (4.2.1) bis (4.2.5) des Korollars.

Zu (4.2.1):

$$\|G_i M_i x_i^*\| = \|G_i(M_i x_i^* - K_i x_i^*)\| \le \|G_i\| \|(H_i L - H_i L H_i) N x_i^*\| .$$

Hier dürfen wir nun nicht einfach durch $\|H_i L D_i\| m$ abschätzen, da bei den von uns angestrebten Anwendungen $\|H_i L D_i\| \to 0$ im allgemeinen nicht zu erreichen ist. Mit Hilfe von Hauptsatz (4.4) können wir jedoch besser abschätzen:

$$\|H_i L D_i N x_i^*\| \le \|H_i L D_i (N x_i^* - N x^*)\| + \|H_i L D_i N x^*\| ;$$

(4.4.5) liefert $\|N x_i^* - N x^*\| \le m' \|x_i^* - x^*\| \le m' A' \|K x^* - K_i x^*\| \le$

$\le m' A \{\|D_i L\| m + \|D_i f\|\}$, wobei A eine aus (4.4.5) ersichtliche Konstante bedeutet. Insgesamt haben wir also

$$\|G_i M_i x_i^*\| \le B m' A \|H_i L D_i\| \{m \|D_i L\| + \|D_i f\|\} + B \|H_i L D_i N x^*\| =: r_i .$$

Zu (4.2.2a):

$$\|G_i M_i'(x_i^*) - I\| \le B \|M_i'(x_i^*) - K_i'(x_i^*)\| = B \|H_i L D_i N'(x_i^*)\|$$

$$\le B m \|A\| H_i L D_i \| \{m \|D_i L\| + \|D_i f\|\} + B \|H_i L D_i N'(x^*)\| =: d_i .$$

Da $\|M_i'(x_i^*) G_i - I\| = \|M_i'(x_i^*) G_i - K_i'(x_i^*) G_i\| \le B \|H_i L D_i N'(x_i^*)\|$,

gilt auch (4.2.2b).

Zu (4.2.3):

Wählt man $r_i \le r - \bar{r}$, so gilt $\bar{S}(x_i^*, r_i) \subset \bar{S}(x_0, r)$, also gilt für $x \in \bar{S}(x_i^*, r_i)$ die Abschätzung

$$\|G_i M_i'(x)\| = \|G_i H_i L H_i N''(x)\| \le B \|L\| \|H_i\|^2 m =: k_i .$$

Wegen (6.1.2a) und (6.1.4) gilt nun $\lim_{i \to \infty} d_i = 0$. Mit

$h_i := \dfrac{k_i r_i}{(1-d_i)^2}$ erhalten wir aus (6.1.2b) und (6.1.3b) $\lim_{i \to \infty} h_i = 0$.

Mit $r_i^{\pm} := \frac{1\pm\sqrt{1-h_i}}{h_i}\,\frac{\eta_i}{1-d_i}$ und $r_i := \frac{1}{2}(r_i^+ + r_i^-)$ sind auch (4.2.4) und (4.2.5) des Korollars erfüllt, da $\frac{1}{2}(r_i^+ + r_i^-) = \frac{2(1-d_i)}{k_i} \to 0$, falls $\|H_i\| \to \infty$. Also folgt die Existenz und Eindeutigkeit einer Lösung x_i^0.

Verzichtet man auf $\|H_i\| \to \infty$, folgt immer noch die Existenz, da für jedes beliebige $r_i := r - r^- - \varepsilon > 0$ für alle i stets gilt

$$r_i^- < r_i \quad \text{für } i \text{ genügend groß .}$$

Denn es ist $\lim\limits_{i\to\infty} \frac{1-\sqrt{1-h_i}}{h_i} = 1$, also $\lim\limits_{i\to\infty} r_i^- = 0$, da wegen (6.1.2a) und (6.1.3a) $\lim\limits_{i\to\infty} \eta_i = 0$.

Nach Korollar (4.2) haben wir überdies die Abschätzung $\|x_i^0 - x_i^*\| \le r_i^-$, also die Konvergenzbehauptung.

(6.2) Bemerkung

(6.1.2b), (6.1.3b) und (6.1.5) implizieren (6.1.2a) und (6.1.3a).

Wie in Satz (5.2) wollen wir nun auch die Gleichung $M_i x_i = 0$ in E_i vermöge des Isomorphismus $h_i : E_i \to F_i$ in den Teilraum F_i transportieren. Wir betrachten also in F_i eine Näherungsgleichung

$$N_i y_i \equiv y_i - h_i H_i L H_i N h_i^{-1} y_i - h_i H_i f \,. \tag{6.3}$$

Wenn wir Satz (5.2) und Satz (6.1) geeignet kombinieren, erhalten wir eine Konvergenzaussage für die Lösungen der Gleichung (6.3). Diese Gleichung wird in unseren späteren Anwendungen auch die praktisch zu lösende Gleichung sein.

Um das Ergebnis einfacher formulieren zu können, wollen wir uns hier auf den schon in Bemerkung (5.3) erwähnten Fall beschränken, daß $H_i : E \to E_i$ eine Projektion ist.

(6.4) SATZ

Es seien die Voraussetzungen von Satz (6.1) erfüllt, außerdem für $i \to \infty$

$$(6.4.2) \quad \begin{cases} \|H_i L D_i\| \|D_i f\| \|h_i\| \to 0 \\ \|H_i L D_i\| \|D_i L\| \|h_i\| \to 0 \end{cases}$$

$$(6.4.3) \quad \|H_i L D_i N x^*\| \|h_i\| \to 0 .$$

Dann gibt es ein $i_o \in \mathbb{N}$, so daß für $i \geq i_o$ die Gleichung (6.3) in F_i eine Lösung y_i^o hat. Ist y_i^* Lösung von $L_i y_i = 0$, so gilt die Konvergenz

$$\|y_i^o - y_i^*\| \to 0 .$$

Gilt überdies $\|H_i\| \to \infty$ für $i \to \infty$, so ist y_i^o lokal eindeutig.

Beweis:

Wie beim Beweis von Satz (5.2) sieht man, daß (6.3) in $\overline{S}(h_i x_i^*, \|h_i\| r_i)$ eine Lösung $y_i^o = h_i x_i^o$ hat, die in $\overline{S}(h_i x_i^*, \|h_i^{-1}\|^{-1} r_i)$ eindeutig ist, falls x_i^o in $\overline{S}(x_i^*, r_i)$ eindeutig ist.

Wegen $\|y_i^o - y_i^*\| \leq \|h_i\| \|x_i^o - x_i^*\| \leq \|h_i\| r_i^-$, läßt sich mit Hilfe von (6.4.2) und (6.4.3) wie im Beweis von Satz (6.1) zeigen, daß $\lim_{i \to \infty} \|h_i\| r_i^- = 0$

DRITTEL KAPITEL

EIN VERFAHREN ZUR PRAKTISCHEN BESTIMMUNG DER LÖSUNG EINER HAMMERSTEINSCHEN INTEGRALGLEICHUNG

§ 7. Definition und Konvergenz des Ersatzsystems

Als eine Anwendung der im zweiten Kapitel entwickelten Theorie wollen wir nun ein Kollokationsverfahren zur Lösung einer Hammersteinschen Integralgleichung

$$(7.1) \qquad x(s) - \int_{-1}^{1} L(s,t)N(t,x(t))dt - f(s) = 0 \ , \quad s \in I := [-1,1]$$

mit Hilfe von Satz (6.4) angeben. Zunächst führen wir Bezeichnungen ein und erklären das Vorgehen.

Zu einer Stützstellenmatrix (s_{ni}) mit $-1 \le s_{ni} < s_{n,i+1} \le 1$, $i \in \{0,1,\dots,n\}$, $n \in \mathbb{N}$ sei

$$w_n(s) := \prod_{i=0}^{n} (s - s_{ni})$$

$$l_{ni}(s) := \frac{w_n(s)}{(s-s_{ni})w_n'(s_{ni})}, \quad \lambda_n(s) := \sum_{i=0}^{n} |l_{ni}(s)|,$$

$$\lambda_n := \max_{s \in I} \lambda_n(s) \ .$$

Mit $L_n[f;s] := \sum_{i=0}^{n} f(s_{ni})\, l_{ni}(s)$ bezeichnen wir das Lagrange-Interpolationspolynom einer Funktion $f \in C(I)$.

Mit $\omega_f(\delta) := \sup_{|s-s'| \le \delta} |f(s)-f(s')|$ bezeichnen wir den Stetigkeitsmodul einer Funktion. $\mathrm{Lip}_M \alpha$ sei wie üblich die Klasse aller Funktionen, die in $[-1,1]$ einer Lipschitzbedingung $|f(s)-f(s')| \le M|s-s'|^{\alpha}$ genügen. Wir erinnern an folgende Tatsachen:

(7.2) $\qquad f \in \mathrm{Lip}_M\alpha \Leftrightarrow \omega_f(\delta) \le M\delta^\alpha \ .$

Ist $f \in C(I)$ und P_n die Menge der Polynome vom Höchstgrad n, so bezeichnen wir mit

$$E_n(f) := \inf_{p\in P_n} \|f-p\|$$

die algebraische Approximationskonstante von f. Es gilt:

(7.3) $$\lim_{n\to\infty} \lambda_n E_n(f) = 0 \Rightarrow \lim_{n\to\infty} \sup_{s\in I} |L_n[f;s]-f(s)| \le$$
$$\le (1+\lambda_n)E_n(f) = 0 \ .$$

Satz von Jackson:

(7.4) Für $f \in C(I)$ gilt $E_n(f) \le c\ \omega_f(n^{-1})$.

Für die Tschebyscheffknoten $s_{ni} := \cos\frac{i-}{n}$ gilt

(7.5) $$\lambda_n \le A + B \ln n \ .$$

Zum Beweis sei auf NATANSON [1965] und MEINARDUS [1964] verwiesen.

Bekanntlich gilt für jede Stützstellenmatrix $\lambda_n \ge C \ln n$, so daß die Tschebyscheffknoten für die hinreichende Bedingung (7.3) optimal sind. Zwei weitere Gründe machen die T-Knoten als Kollokationspunkte besonders geeignet. Erstens genügen die Tschebyscheffpolynome bezüglich dieser Knoten einer Summenorthogonalität, die sich für numerische Zwecke als günstig erweist, zweitens sind sie etwa den Nullstellen des T-Polynoms vom Grade $n+1$, die auch (7.5) genügen, vorzuziehen, da sich der Interpolationsfehler im allgemeinen besser abschätzen läßt, wie ELLIOTT [1964] zeigt.

Mit dieser Stützstellenmatrix definieren wir uns nun folgendermaßen geeignete Räume und Operatoren.

Sei $E := C(I)$ versehen mit der T-Norm, E_n der Raum der Polynome vom Höchstgrad n und $F_n := \mathbb{R}^{n+1}$ ebenfalls mit der T-Norm versehen. H_n sei derjenige Operator, der jeder Funktion $f \in E$ ihr eindeutig bestimmtes Interpolationspolynom vom Höchstgrad n zuordnet. h_n möge ein Element $p \in E$ in den Vektor der Stützwerte $(p(s_{no}), p(s_{n1}),\ldots,p(s_{nn}))' \in F_n$ abbilden. Aus (7.5) folgt dann

$$\|H_n\| \leq A + B \ln n \,. \tag{7.6}$$

Leicht sieht man, daß $\|h_n\| = 1$.

Sei nun $L_n[x;s] = \sum_{k=0}^{n} a_k l_k(s)$ das Interpolationspolynom der Lösungsfunktion $x(s)$, wobei $a_k := x(s_k)$ und $l_k(s) := l_{nk}(s)$, so ist $y_n := (a_o,\ldots,a_n)' \in F_n$. Der Reihe nach erhalten wir

$$h_n^{-1}y_n : h_n^{-1}y_n(t) = \sum_{k=0}^{n} a_k l_k(t)$$

$$Nh_n^{-1}y_n : Nh_n^{-1}y_n(t) = N(t, \sum_{k=0}^{n} a_k l_k(t))$$

$$H_nNh_n^{-1}y_n : H_nNh_n^{-1}y_n(t) = L_n[Nh_n^{-1}y_n;t] = \sum_{k=0}^{n} N(s_k,a_k)l_k(t)$$

$$LH_nNh_n^{-1}y_n : LH_nNh_n^{-1}y_n(s) = \int_{-1}^{1} L(s,t)H_nNh_n^{-1}y_n(t)dt =$$

$$= \sum_{k=0}^{n} N(s_k,a_k) \int_{-1}^{+1} L(s,t)l_k(t)dt$$

und damit schließlich mit der Bezeichnung wie in Gleichung (6.3)

$$N_n y_n = (a_i - \sum_{k=0}^{n} \int_{-1}^{+1} L(s_i,t)l_k(t)dt\, N(s_k,a_k) - f(s_i))_{i=0,\ldots,n} \,,$$

also das nichtlineare Gleichungssystem für den Vektor $(a_0,\ldots,a_n)'$

$$(7.7) \qquad a_i - \sum_{k=o}^{n} N(s_k,a_k) \int_{-1}^{+1} L(s_i,t)l_k(t)dt - f(s_i) = 0, \qquad i=0,\ldots,n;$$

hierbei haben wir $s_i := s_{ni}$ $(i=0,\ldots,n)$ gesetzt.

Mit den stetigen Funktionen $L(s,t)$ und $N(t,u)$, also $L(s,t) \in C(I\times I)$ und $N(t,u) \in C(I\times\mathbb{R})$ sind die Operatoren $L,N : E \to E$ folgendermaßen erklärt:

$$Lx = y, \quad y(s) = \int_{-1}^{1} L(s,t)x(t)dt,$$

$$Nx = y, \quad y(t) = N(t,x(t)) .$$

Gleichung (7.1) ist also äquivalent mit

$$(7.8) \qquad Kx \equiv x - LNx - f = 0 .$$

Einen Satz (4.1) entsprechenden Existenzsatz zitieren wir aus KANTOROWITSCH [1964], S. 589.

(7.9) SATZ

Ist $x_0(s)$ eine Näherungslösung für (7.1) und sind die folgenden Bedingungen erfüllt:

(7.9.1) Für den Kern $\bar{L}(s,t) := L(s,t)N'_u(t,x_0(t))$ hat die Integralgleichung

$$\bar{x}(s) - \int_{-1}^{1} \bar{L}(s,t)\bar{x}(t)dt - \varepsilon_0(s) = 0 ,$$

wobei

$$\varepsilon_0(s) = x_0(s) - \int_{-1}^{1} L(s,t)N(t,x_0(t))dt - f(s)$$

eine Resolvente $G_{x_0}(s,t)$ mit

$$\max_{s\in I} \int_{-1}^{1} |G_{x_0}(s,t)|\,dt \le B;$$

(7.9.2) $\quad \|\varepsilon_0\| \le \eta$;

(7.9.3) Für $(s,u) \in \bar{S} := \{(s,u) \mid s\in I,\ \|u-x_0\| \le 2(1+B)\eta\}$

gilt

$$\int_{-1}^{1} |L(s,t)N''_{u^2}(t,u)|\,dt \le M .$$

(7.9.4) $\quad h := (1+B)^2 M\eta < \frac{1}{2}$.

Dann gilt mit $r^{\pm} := \frac{1 \pm \sqrt{1-2h}}{h}(1+B)\eta$ die Aussage von Satz (4.1).

Beweis:

Wir zeigen, daß mit $B_0 := 1 + B$ $\eta_0 := (1+B)\eta$ und M die Voraussetzungen von Satz (4.1) erfüllt sind.

Es ist $\|K'(x_0)^{-1}\| \le 1 + \|G\| \le 1 + B$ wegen Voraussetzung (7.9.1), $\|x_1-x_0\| \le B_0\|\varepsilon_0\| \le (1+B)\eta$ wegen Voraussetzung (7.9.1) und (7.9.2), $\|K''(x)\| \le M$ wegen (7.9.3). Daß $r^- < 2(1+B)\eta < r^+$, rechnet man leicht nach.

Dieser Beweis zeigt also, daß die Voraussetzungen von Satz (7.9) die Voraussetzungen von Satz (4.1) für den gemäß (7.8) definierten Operator implizieren.

Nach diesen Vorbereitungen können wir nun einen Konvergenzsatz für die Lösungen des Systems (7.7) formulieren.

(7.10) SATZ

Für die Hammersteinsche Integralgleichung

$$x(s) - \int_{-1}^{+1} L(s,t)N(t,x(t))dt - f(s) = 0$$

seien mit der Anfangsnäherung $x_0(s)$ die Voraussetzungen von Satz (7.9) erfüllt. Ferner gelte:

(7.10.1) $\quad f \in \mathrm{Lip}\,\alpha\ ,\ x_0 \in \mathrm{Lip}\,\alpha$

(7.10.2) $$\bigvee_{M} \bigwedge_{t\in I} \bigwedge_{s,s'\in I} |L(s,t)-L(s',t)| \le M|s-s'|^{\alpha}$$

(7.10.3a) $$\bigvee_{M_1} \bigwedge_{u\in\bar{S}} \bigwedge_{t,t'\in I} |N(t,u)-N(t',u)| \le M_1|t-t'|^{\alpha}$$

(7.10.3b) $$\bigvee_{M_2} \bigwedge_{u\in\bar{S}} \bigwedge_{t,t'\in I} |N'_u(t,u)-N'_u(t',u)| \le M_2|t-t'|^{\alpha}$$

(7.10.4) (s_{ni}) $(i=0,\ldots,n;\ n=1,2,\ldots)$ ist eine Stützstellenmatrix, für deren Lebesguefunktion gilt

$$\lambda_n \le A + B \ln n\ .$$

Dann gibt es ein $n_0 \in \mathbb{N}$, so daß für $n \ge n_0$ das Gleichungssystem

$$a_{ni} - \sum_{k=0}^{n} N(s_{nk},a_{nk}) \int_{-1}^{1} L(s_{ni},t)l_{nk}(t)dt - f(s_{ni}) = 0 \quad (i=0,\ldots,n)$$

lokal eindeutig lösbar ist, und es gilt für die Lösung $\underline{a}_n^* := (a_{no}^*,\ldots,a_{nn}^*)'$

$$\|\underline{a}_n^* - \underline{x}^*\| \to 0,\quad n \to \infty\ .$$

Hierbei ist $\underline{x}^* := (x^*(s_{no}),\ldots,x^*(s_{nn}))'$, wenn mit $x^*(s)$ die eindeutig bestimmte Lösung von Gleichung (7.1) bezeichnet wird.

Wir verifizieren zunächst die Voraussetzungen von Satz (6.1). Hierzu zeigen wir zuerst gemäß Bemerkung (6.2) die Voraussetzungen (6.1.1), (6.1.2b), (6.1.3b) und (6.1.5) von Satz (6.1).

Zu (6.1.1):

$$x_o \in \mathrm{Lip}\,\alpha \Rightarrow E_n(x_o) \le c\, \omega_{x_o}(\tfrac{1}{n}) \le \frac{cM}{n^\alpha} \Rightarrow \lambda_n E_n(x_o) \le$$

$$\le (A+B \ln n)\frac{cM}{n^\alpha} ,$$

also nach (7.3) $\quad \|D_n x_o\| \to 0$.

Analog schließt man für f und L, letzteres wegen

$$D_n L \le \max_{s\in I} \int_{-1}^{1} |L(s,t)-L_n[L;s]|dt \quad \text{und Voraussetzung (7.10.2).}$$

Zu (6.1.2b):

Es ist $\|H_n L D_n\| \le \|H_n\|\,\|L\|(1+\|H_n\|) = O((\ln n)^2)$, also

$$\|H_n L D_n\|\,\|D_n f\|\,\|H_n\|^2 = O(\frac{(\ln n)^4}{n^\alpha}) .$$

Zu (6.1.3b):

Ist $x^*(s)$ Lösung, so gilt $x^*(s) = \int_{-1}^{1} L(s,t)N(t,x^*(t))dt + f(s)$.

Wegen $\left|\int_{-1}^{1} [L(s,t)-L(s',t)]N(t,x^*(t))dt\right| \le$

$$\le 2M|s-s'|^\alpha \int_{-1}^{1} |N(t,x^*(t))|dt \quad \text{und}$$

$f \in \mathrm{Lip}\,\alpha$, ist auch $x^* \in \mathrm{Lip}\,\alpha$. Nun gilt wegen der stetigen Differenzierbarkeit von $N(t,u)$ als Funktion von u und wegen Voraussetzung (7.10.3a) mit geeigneten Konstanten a, a_1

$|N(t,x^*(t))-N(t',x^*(t'))| \le |N(t,x^*(t))-N(t,x^*(t'))| +$

$+ |N(t,x^*(t')) - N(t',x^*(t'))| \le a|x^*(t)-x^*(t')| +$

$+ M_1|t-t'|^\alpha \le a_1|t-t'| + M_1|t-t'|^\alpha$, also $Nx^* \in \mathrm{Lip}\,\alpha$.

Folglich ist $\|D_n Nx^*\| = O(\frac{\ln n}{n^\alpha})$ und

$$\|H_n L D_n Nx^*\| \|H_n\|^2 = O(\frac{(\ln n)^4}{n^\alpha}) .$$

Zu (6.1.4):

Wegen der zweimaligen stetigen Differenzierbarkeit von $N(t,u)$ als Funktion von u und wegen Voraussetzung (7.10.3b) hat man mit geeigneten Konstanten b, b_1

$|N'_u(t,x^*(t))-N'_u(t',x^*(t'))| \le b|x^*(t)-x^*(t')| +$

$+ M_2|t-t'|^\alpha \le (b_1+M_2)|t-t'|^\alpha$, also $N'_u(x^*) \in \mathrm{Lip}\,\alpha$.

Da $\lim_{n\to\infty} \frac{(\ln n)^4}{n^\alpha} = 0$ für jedes $\alpha > 0$, sind die Voraussetzungen von Satz (6.1) verifiziert. Wegen $\|h_i\| = 1$ folgen (6.4.2) und (6.4.3) von Satz (6.4) trivial, also ist Satz (6.4) anwendbar und damit Satz (7.10) bewiesen.

§ 8. Abschätzung des Diskretisierungsfehlers

MYSOVSKIH [1962] gibt Fehlerabschätzungen für Näherungssysteme an, die durch Quadratur des Integrals und anschließende Diskretisierung - also die übliche Kollokationsmethode für Integralgleichungen - entstanden sind. Unser Näherungssystem kann man sich auch dadurch entstanden denken, daß auf das Integral eine verallgemeinerte Quadratur angewendet und anschließend diskretisiert wird. Die verallgemeinerte Quadratur wurde von ATKINSON [1966] zur Lösung linearer Integralgleichungen angewendet. Dort konnten auch Kerne mit absolut integrierbaren Singularitäten zugelassen werden. Es scheint mög-

lich, unser Verfahren für nichtlineare Hammersteinsche Integralgleichungen auch dahingehend zu erweitern. Für die hier betrachteten Gleichungen ist es jedoch ausreichend, die Ergebnisse von MYSOVSKIH [1962] auf verallgemeinerte Quadraturen zu übertragen. Zunächst geben wir einige Fehlerabschätzungen für verallgemeinerte Interpolationsquadraturen an, wobei wir wieder die Stützstellenmatrix der T-Knoten zugrunde legen.

FRASER-WILSON [1966] geben für den Interpolationsfehler

$$R_n(f;t) := f(t) - L_n(f;t)$$

einer genügend oft differenzierbaren Funktion $f \in C^{n+1}(I)$ die Abschätzung

$$(8.1) \quad |R_n(f;t)| \le \frac{(1-t^2)|\overline{U}_{n-1}(t)|}{(n+1)!}\|f^{(n+1)}\| \le$$

$$\le \frac{n}{2^{n-1}(n+1)!}\|f^{(n+1)}\| , \quad t \in I$$

an; hierbei bezeichnet $\overline{U}_n(t)$ das auf Höchstkoeffizient 1 normierte Tschebyscheffsche Polynom zweiter Art. Betrachten wir den Fehler bei "verallgemeinerter Interpolation" für festes $g \in C(I)$

$$V_n(f;t) := g(t)f(t) - g(t)L_n(f;t) ,$$

so gilt

$$\|V_n(f;\cdot)\| \le \|g\| \, \|R_n(f;\cdot)\| , \quad f, g \in C(I) .$$

Für den Quadraturfehler $R_n(f) = \int_{-1}^{1} R_n(f;t)dt$ erhält man aus (8.1) die Abschätzung

$$|R_n(f)| \leq \frac{\|f^{(n+1)}\|}{2^{n-1}(n+1)!} \cdot \tag{8.2}$$

Überträgt man das Vorgehen von FRASER-WILSON auf die verallgemeinerte Interpolationsquadratur

$$\int_{-1}^{1} g(t)f(t)dt = \int_{-1}^{1} g(t)L_n(f;t)dt + \tag{8.3}$$

$$+ \int_{-1}^{1} g(t)R_n(f;t)dt \ ,$$

so erhält man eine Abschätzung für den Quadraturfehler

$$V_n(f) := \int_{-1}^{1} V_n(f;t)dt$$

$$|V_n(f)| \leq \frac{\|g\| \, \|f^{(n+1)}\|}{2^{n-1}(n+1)!} \ , \quad n \geq 2 \ .$$

LOCHER [1969] konnte die Abschätzung (8.2) für gerades n und $f \in C^{n+2}(I)$ verbessern:

$$|R_n(f)| \leq \frac{\|f^{(n+2)}\|}{2^{n-2}n(n+2)!} \ .$$

Für eine Funktion $g(t)$, die in $[-1,1]$ ihr Vorzeichen nicht wechselt, kann man den Beweis von LOCHER modifizieren und erhält eine Abschätzung

$$|V_n(f)| \leq \frac{\|g''\| \, \|f^{(n+2)}\|}{2^{n-2}n(n+2)!} \ . \tag{8.4}$$

Steht die Ableitung nicht zur Verfügung, kann man eine Abschätzung mit Hilfe der algebraischen Approximationskonstanten von f angeben (LOCHER-ZELLER [1968]):

$$(8.5) \qquad |R_n(f)| \le \{2+\sum_{k=0}^{n}|a_k|\}E_n(f) \le 4\,E_n(f)\,.$$

$$a_k := \int_{-1}^{+1} l_k(t)dt\,.$$

Wählt man in der verallgemeinerten Quadraturformel (8.3) für festes g die Bezeichnung

$$\int_{-1}^{1} g(t)f(t)dt = \sum_{k=0}^{n} A_k\, f(t_k) + V_n(f)\,,$$

$$A_k := \int_{-1}^{+1} f(x)l_k(x)dx\,,$$

so findet man leicht eine (8.5) entsprechende Abschätzung

$$(8.6) \qquad |V_n(f)| \le \{\int_{-1}^{1}|g(t)|dt + \sum_{k=0}^{n}|A_k|\}E_n(f) \le$$

$$\le \{2\,\|g\| + \sum_{k=0}^{n}|A_k|\}E_n(f)\,.$$

$E_n(f)$ kann man z.B. mit einer Methode von EHLICH-ZELLER [1964] numerisch gut erfassen. Eine Fehlerabschätzung der Größenordnung des Quadraturfehlers, die die zur Quadratur benötigten Funktionswerte $f(t_k)$ $(k=0,\ldots,n)$ verwendet, wird z.B. von LOCHER [1969] angegeben

$$|R_n(f)| \le \frac{4}{n^2-1}\,|L_{n-1}(f)| + \varepsilon\,,$$

wo ε klein aber unbekannt ist; $L_n(f)$ ist die diskrete Approximationskonstante bezüglich der T-Knoten. Wir wollen hier darauf verzichten, genaue Untersuchungen über mögliche Fehlerabschätzungen bei verallgemeinerter Quadratur durchzuführen, da wir lediglich eine praktikable Schranke bei der

numerischen Durchführung unseres Verfahrens zur Lösung von Integralgleichungen benötigen. Es sei noch bemerkt, daß sich die von uns verwendete Schranke (8.6) in $C[-1,1]$ nicht verbessern läßt (vgl. LOCHER [1968], S. 10).

Zur Abschätzung von $E_n(f)$ geht man nun so vor, daß zunächst f durch ein genügend genaues Polynom $P_m(t)$ vom Grade m ersetzt wird. Dann gilt

(8.7) $$E_n(f) \le \varepsilon_m + \|D_m(f;\cdot)\| \quad , \quad \text{wobei}$$

$$D_m(f;t) := P_m(t) - L_n(f;t) \quad \text{und} \quad \varepsilon_m := \|f-P_m\|$$

der Interpolationsfehler ist.

Die weitere Abschätzung erfolgt mit dem nächsten Satz.

(8.8) SATZ (EHLICH-ZELLER [1964])

Aus

(8.8.1) $$|D_m(f;t_k)| \le \varepsilon \quad (t_k=\cos\tfrac{k\pi}{lm};\ k=0,\dots,lm;\ l\ge 2)$$

folgt

(8.8.2) $$\|D_m(f;\cdot)\| \le \varepsilon(\cos\tfrac{\pi}{2l})^{-1} \ ;$$

diese Abschätzung ist scharf.

Für $m-n \le 4$ beweist CLENSHAW [1964], daß $L_n(f;t)$ bezüglich der T-Knoten höchstens um den Faktor 2 schlechter ist als die beste Polynomapproximation vom Höchstgrade n an $P_m(t)$. Für $m-n = 5$ ist diese Behauptung numerisch nachgeprüft, für $m-n > 5$ noch unbewiesen. Man darf also erwarten, daß die Abschätzung (8.7) nicht zu schlecht ist.

Für eine feste Funktion $K(s,t)$ mit $K \in C(I\times I)$ betrachten wir nun für $s \in I$ die verallgemeinerte Quadratur

$$(8.9)\qquad \int_{-1}^{1} K(s,t)x(t)dt = \sum_{k=0}^{n} x(t_k) \int_{-1}^{1} K(s,t)l_k(t)dt + \\ + V_n(x;s) .$$

Für festes K und n schreiben wir

$$A_k(s) := \int_{-1}^{1} K(s,t)l_k(t)dt .$$

Eine grundlegende Bedeutung für unsere Fehlerabschätzung hat die Funktion

$$(8.10)\qquad \varepsilon(s,t) := \int_{-1}^{1} K(s,\tau)K(\tau,t)d\tau - \sum_{k=0}^{n} A_k(s)K(t_k,t) .$$

Sie stellt den Quadraturfehler des Kernes $K(s,t)$ als Funktion von s unter der Quadratur (8.9) dar. Eine bei gewöhnlicher Quadratur (8.10) entsprechende Funktion wurde von MYSOVKIH [1956] eingeführt.

Mit (8.10) definieren wir nun $(s_i = t_i,\ i=0,\dots,n)$

$$(8.11)\qquad \begin{cases} \varepsilon := \max\limits_{0\le i\le n} \int_{-1}^{1} |\varepsilon(s_i,t)|dt \\ \|\varepsilon\| := \max\limits_{-1\le s\le 1} \int_{-1}^{1} |\varepsilon(s,t)|dt . \end{cases}$$

$\|\varepsilon\|$ ist die T-Norm des linearen Integraloperators, der durch den Kern $\varepsilon(s,t)$ definiert wird.

Für den Kern $K(s,t)$ definieren wir

$$(8.12)\qquad \begin{cases} K = \max\limits_{0\le i\le n} \int_{-1}^{1} |K(s_i,t)|dt \\ \|K\| := \max\limits_{-1\le s\le 1} \int_{-1}^{1} |K(s,t)|dt \\ K_1 := \max\limits_{-1\le s\le 1} \sum_{k=0}^{n} |A_k(s)| . \end{cases}$$

Es gelten die Abschätzungen $\varepsilon \leq \varepsilon''$ und $K \leq \|K\|$.

Das Näherungssystem (7.7) ist eine andere Schreibweise für die Gleichung

$$N_n a_n = 0 \quad ,$$

wobei N_n der in (6.3) definierte Operator des $\mathbb{R}^{n+1}$ in sich ist. Mit den erklärten Bezeichnungen können wir nun die Abschätzung des Diskretisierungsfehlers formulieren.

(8.13) SATZ

Sei $\underline{a}^* = (a_0^*,\ldots,a_n^*)'$ eine Lösung des Systems (7.7). $\tilde{x}(s)$ sei eine in I stetige Funktion mit $\tilde{x}(s_i) = a_i^*$ $(i=0,\ldots,n)$, und es sind die folgenden Bedingungen erfüllt:

(8.13.1) $\|[N_n'(\underline{a}^*)]^{-1}\| \leq P$;

(8.13.2) Für den Kern $K(s,t) := L(s,t)N_u'(t,\tilde{x}(t))$ und die durch (8.12), (8.13) definierten Größen gilt

$$q := \|\varepsilon\| + K_1 P\varepsilon < 1 \quad ;$$

(8.13.3) Das Residuum $\tilde{\varepsilon}(s) := \tilde{x}(s) - \int_{-1}^{1} L(s,t)N(t,\tilde{x}(t))dt - f(s)$ bezüglich $\tilde{x}(s)$ genügt der Ungleichung $\|\tilde{\varepsilon}\| \leq \varepsilon$;

(8.13.4) Für $s,t \in I$ und $\|u-\tilde{x}\| \leq 2(1+B)\varepsilon$ mit

$$B := \frac{\|K\|+K_1 P K+q}{1-q}$$

sind $L(s,t)N(t,u)$, $L(s,t)N_u'(t,u)$ und $L(s,t)N_{u^2}''(t,u)$ stetig und es gilt

$$\int_{-1}^{1} |L(s,t)N_{u^2}''(t,u)|dt \leq M \quad ;$$

(8.13.5) $$h := (1+B)^2 M_n < \frac{1}{2} .$$

Dann gilt die Fehlerabschätzung

$$\|\tilde{x}-x^*\| \leq \frac{1-\sqrt{1-2h}}{h} (1+B)\eta .$$

Beweis:

Der Satz folgt aus Satz (7.9), wenn man zeigt, daß für die Resolvente $G_{\tilde{x}}(s,t)$ der linearen Integralgleichung

$$\bar{x}(s) - \int_{-1}^{1} K(s,t)\bar{x}(t)dt - f(s) = 0$$

gilt

$$\max_{s\in I} \int_{-1}^{1} |G_{\tilde{x}}(s,t)|dt \leq \frac{\|K\|+K_1 PK+q}{1-q} .$$

Diese Ungleichung hat MYSOVSKIH [1962] in dem dortigen Theorem 1 bei gewöhnlicher Quadratur bewiesen. Der Beweis läßt sich ohne Schwierigkeiten auf den Fall verallgemeinerter Quadratur übertragen. Daher wollen wir an dieser Stelle auf den Beweis verzichten.

§ 9. Berechnung der Gewichte bei verallgemeinerter Quadratur

Die numerische Behandlung des Gleichungssystems (7.7) besteht aus zwei wesentlichen Aufgaben:

1) Berechnung der Größen $A_{ik} := \int_{-1}^{1} L(s_i,t)l_k(t)dt$, $i,k=0,\ldots,n$;

2) Lösung des nichtlinearen Gleichungssystems durch Newton-Iteration mit Erfassung der Rundefehler.

In diesem Paragraphen wollen wir Aufgabe 1) betrachten, wobei wir im Hinblick auf unsere numerischen Beispiele in § 11 auf das Intervall $[0,1]$ transformieren. Die Stützstellen seien also $s_i := \frac{1}{2}(1+\cos\frac{i\pi}{n})$ $(i=0,\dots,n)$; mit $T_i^*(t)$ $(i=0,1,\dots)$ bezeichnen wir die auf das Intervall $[0,1]$ transformierten Tschebyscheffpolynome erster Art:

$$T_i^*(t) := T_i(2t-1) , \quad t \in [0,1] .$$

Hierbei ist $T_i(t)$ $(i=0,1,\dots)$ das gewöhnliche T-Polynom. Wegen der Summenorthogonalität

$$\sum_{k=0}^{n}{}'' \cos\frac{r\pi k}{n} \cos\frac{s\pi k}{n} = \begin{cases} 0 , & r \neq s \\ n , & r = s = 0,n \\ \frac{n}{2} , & \text{sonst} \end{cases} \qquad (0\le r,s\le n)$$

$(\sum_{k=0}^{n}{}'' a_k := \frac{1}{2} a_0 + \frac{1}{2} a_n + \sum_{k=1}^{n-1} a_k)$, hat das Lagrange-Grundpolynom die Form

$$l_{nk}(t) = l_k(t) = \begin{cases} \frac{1}{n} \cdot \sum_{j=0}^{n}{}'' \cos\frac{k\pi j}{n} T_j^*(t) , & \text{falls } k=0 \text{ oder } k=n \\ \frac{2}{n} \cdot \sum_{j=0}^{n}{}'' \cos\frac{k\pi j}{n} T_j^*(t) , & \text{sonst} \end{cases} .$$

In vielen Fällen ist es ausreichend, $L(s_i,t)$ durch eine endliche T-Reihe mit wenigen Termen zu approximieren:

$$L(s_i,t) \sim \sum_{r=0}^{m} B_{ir} T_r^*(t) .$$

Die dann in A_{ik} auftretenden Integrale $\int_0^1 T_r^*(t) l_k(t) dt$ kann man nach einiger elementarer Rechnung in der Form

$$\int_0^1 T_r^*(t)l_k(t)dt = \frac{-1}{2n}\Big[\sum_{s=0}^{n}{}'' \frac{1+(-1)^{r+s}}{(r+s)^2-1} T_s^*(s_k) +$$

$$+ \sum_{s=0}^{r-2}{}' \frac{1+(-1)^{r-s}}{(r-s)^2-1} T_s^*(s_k) +$$

$$+ \sum_{s=r+2}^{n}{}' \frac{1+(-1)^{r-s}}{(r-s)^2-1} T_s^*(s_k) - 2\cos\frac{r\pi k}{n}\Big]$$

darstellen ($\sum_{k=0}^{n}{}' a_k := \frac{1}{2}a_0 + \sum_{k=1}^{n} a_k$). Die angegebene Darstellung gilt nur für $2 \le r \le n-2$; für andere Werte von r kann man eine entsprechende Darstellung angeben. Für die rechtsstehenden Summen gibt es einen besonders vorteilhaften Algorithmus, der die Rekursionsformel der T-Polynome ausnutzt (vgl. SMITH [1965]).

Eine besonders wichtige Klasse Hammersteinscher Integralgleichungen erfaßt man wegen der Äquivalenz zwischen einer geeigneten Integralgleichung und der Randwertaufgabe

$$y''(t) + f(t,y(t),y'(t)) = 0$$

$$y(0) = y(1) = 0 ,$$

wenn man die Greensche Funktion

$$L(s,t) = \begin{cases} s(1-t) & s \le t \\ t(1-s) & t \le s \end{cases}$$

als Kern wählt. Da $L(s,t)$ als Funktion von t in $[0,1]$ nicht differenzierbar ist, konvergiert die T-Reihe schlecht. In diesem Fall kann man die Integrale

$A_{ik} = \int_0^1 L(s_i,t)l_k(t)dt$ jedoch ausrechnen. Es soll hier nur

das Ergebnis angegeben werden. Mit den Stammfunktionen

$$I_{k1}(t) := \frac{2}{n}\left[\frac{t}{2} + t(t-1)\cos\frac{k\pi}{n} + \frac{1}{4}\sum_{j=2}^{n-1}\left\{\frac{T_{j+1}^*(t)}{j+1} - \frac{T_{j-1}^*(t)}{j-1}\right\}\cos\frac{jk\pi}{n}\right]$$

und

$$I_{k2}(t) := \frac{1}{2n}\left[t^2 + \left\{2t^2-t+\frac{1}{4}\left(\frac{1}{3}T_3^*(t)-2t+1\right)\right\}\cos\frac{k\pi}{n} + \right.$$

$$+ \left\{t^2-t+\frac{1}{2}\left(\frac{1}{3}T_3^*(t)-2t+1\right)\right\} + \frac{1}{8}\left\{\frac{1}{2}T_4^*(t) - T_2^*(t)\right\}\cos\frac{2k\pi}{n} +$$

$$\left. + \frac{1}{4}\sum_{j=3}^{n-1}\left\{\frac{T_{j+2}^*(t)}{j+2} + \frac{2T_{j+1}^*(t)}{j+1} - \frac{2T_{j-1}^*(t)}{j-1} - \frac{T_{j-2}^*(t)}{j-2}\right\}\cos\frac{jk\pi}{n}\right.$$

errechnen wir

$$A_{ik} = I_{k2}(s_i) - s_i I_{k1}(s_i) - (1-s_i)I_{k2}(0) + s_i I_{k1}(1) -$$

$$- s_i I_{k2}(1), \quad i,k=0,\ldots,n .$$

Für $k=0$ oder $k=n$ muß die rechte Seite noch durch 2 dividiert werden. Natürlich kann man die Integrale A_{ik} auch durch numerische Quadratur gewinnen; der Mehraufwand gegenüber der gewöhnlichen Kollokation besteht darin, daß man für jeden Kern $L(s,t)$ eigene Gewichte benötigt, wohingegen bei gewöhnlicher Kollokation die Gewichte

$\int_0^1 l_k(t)\,dt$ für alle Kerne nur einmal zu berechnen sind. Auf längere Sicht betrachtet, fällt dieser Nachteil in der Praxis nicht ins Gewicht.

§ 10. Erfassung der Rundefehler durch Methoden der Intervallarithmetik

Bei der Darstellung der Menge der reellen Zahlen durch die endliche Menge R_M der Maschinenzahlen einer Rechenanlage treten unweigerlich Fehler auf, die sich bei Rechenoperationen fortpflanzen. In unserem Fall muß ein nichtlineares Gleichungssystem (7.7) mittels Newton-Iteration gelöst werden

$$a_i - \sum_{k=0}^{n} A_{ik} N(s_k, a_k) - f(s_i) = 0 \quad i=0,\ldots,n \ .$$

Gewinnt man die Gewichte A_{ik} durch numerische Quadratur, kommt hierdurch eine weitere Fehlergröße dazu. Eine geeignete Methode zur Gewinnung sicherer Schranken für den Lösungsvektor liefert uns die Intervallarithmetik.

Wir definieren den Operator $T : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ durch

$$Tx = (T_0 x, \ldots, T_n x) \ , \quad x = (x_0, x_1, \ldots, x_n)'$$

$$T_i x := x_i - \sum_{k=0}^{n} A_{ik} N(s_k, x_k) - f(s_i) \ , \quad i=0,\ldots,n$$

und lösen die Gleichung $Tx = 0$. Zunächst wollen wir in knapper Form die hier benötigten Begriffe der Intervallarithmetik zusammenstellen.

Unter einer Triplexzahl A verstehen wir ein Tripel von Maschinenzahlen $(\underline{a}, a, \overline{a}) \in R_M^3$ mit der Eigenschaft $\underline{a} \le a \le \overline{a}$ und schreiben

$$A := [\underline{a}, a, \overline{a}] \ .$$

Ist $* \in \{+,-,\times,:\}$ eine der reellen Grundrechenarten, so bezeichnen wir mit $\circledast$ die zugehörige Gleitkommaoperation der

vorgegebenen Maschine. Die Verknüpfung von Triplexzahlen A und B bezeichnen wir mit $\circledast$.

Wir schreiben $x \in A$, falls $\underline{a} \leq x \leq \overline{a}$. Jede Maschinenzahl a kann in natürlicher Weise als Triplexzahl $A := [a,a,a]$ aufgefaßt werden.

Das Ergebnis der Verknüpfung zweier Triplexzahlen A und B ist eine Triplexzahl $C = [\underline{c},c,\overline{c}]$ mit der Eigenschaft

$$c = a * b$$

$$\{z \in R_M : x \in R_M \cap [\underline{a},\overline{a}],\ y \in R_M \cap [\underline{b},\overline{b}] \text{ und } z = x * y\} \subseteq [\underline{c},\overline{c}]$$

Unter einem (n+1)-dimensionalen Triplexvektor C verstehen wir ein (n+1)-Tupel von Triplexzahlen

$$C := (A_0,A_1,\ldots,A_n)'$$

$$A_i := [\underline{a}_i,a_i,\overline{a}_i] , \quad i=0,\ldots,n .$$

Analog definiert man eine Triplexmatrix.

Wir verwenden die übliche Intervallnorm

$$\|A\| := \sup_{x \in A} |x| .$$

Ist $x := (x_0,x_1,\ldots,x_n)$ ein Vektor des $\mathbb{R}^{n+1}$, so schreiben wir

$$x \in C \Leftrightarrow \bigwedge_{0 \leq i \leq n} x_i \in A_i .$$

Für eine Vektornorm $\|x\|$ auf $\mathbb{R}^{n+1}$ erhält man eine Norm für Triplexvektoren

$$\|\mathbf{G}\| := \sup_{\mathbf{x}\in\mathbf{G}} |\mathbf{x}| .$$

Eine Norm für Triplexmatrizen definiert man entsprechend. Wir legen für das folgende die Maximumnorm $\|\mathbf{x}\| := \max_{0\le i\le n} |x_i|$ und die zugeordnete Matrixnorm zugrunde. Es gilt dann für die Norm eines Triplexvektors $\mathbf{G}$:

$$\|\mathbf{G}\| = \sup_{\mathbf{x}\in\mathbf{G}} \|\mathbf{x}\| = \sup_{\mathbf{x}\in\mathbf{G}} \max_{0\le i\le n} |x_i| = \max_{0\le i\le n} \|A_i\| .$$

Unter einer Intervallauswertung $\hat{\mathbf{T}}$ des Operators $\mathbf{T}$ auf dem Triplexvektor $\mathbf{G}$ verstehen wir einen Triplexvektor $\mathbf{B} = (B_0,\ldots,B_n)'$, $B_i = [\underline{b}_i, b_i, \overline{b}_i]$, $i=0,\ldots,n$ mit der Eigenschaft

$$\mathbf{T}(\mathbf{G}) = \bigcup_{\mathbf{x}\in\mathbf{G}} \mathbf{T}\mathbf{x} \subset \mathbf{B} =: \hat{\mathbf{T}}(\mathbf{G})$$

$$b_i = T_i((b_0,\ldots,b_n)') \quad i=0,\ldots,n .$$

Die Intervallauswertung der ersten und zweiten Fréchetableitung definiert man analog.

Zur Fehlererfassung bei intervallmäßiger Ausführung eines Newton-Schrittes zur Lösung des Systems $\mathbf{T}\mathbf{x} = 0$ verwenden wir ein Ergebnis von ROKNE und LANCASTER [1969].

Sei $\mathbf{x}_0$ ein Triplexnäherungsvektor für $\mathbf{T}\mathbf{x} = 0$ und $\hat{\mathbf{G}}_0$ eine Intervallauswertung von $[\hat{\mathbf{T}}'(\mathbf{x}_0)]^{-1}$. Die Newton-Korrektur

$$\mathbf{d} := -\hat{\mathbf{G}}_0 \cdot \hat{\mathbf{T}}(\mathbf{x}_0)$$

ergibt eine verbesserte Näherung

$$\mathbf{x}_1 := \mathbf{x}_0 \oplus \mathbf{d} .$$

Durch den Triplexvektor $\mathbf{X}_1 = (X_{10},\ldots,X_{1n})'$ mit $X_{1i} = [\underline{x}_{1i}, x_{1i}, \overline{x}_{1i}]$ $(i=0,\ldots,n)$ ist ein gewöhnlicher Vektor $x_1 := (x_{10},\ldots,x_{1n})$ definiert. Sei

$r := \max_{0\le i\le n} \max\{|\underline{x}_{1i}-x_{1i}|, |x_{1i}-\overline{x}_{1i}|\}$ und $l_0 \in R_M$ die kleinste berechenbare obere Schranke für $\|\hat{\omega}_0 \otimes \hat{T}''_{K_0}\|$, dann gilt unter diesen Voraussetzungen der folgende Satz.

(10.1) SATZ

Sei $\mathbf{X} = (X_0,X_1,\ldots,X_n)'$, $X_i = (\underline{x}_i, x_i, \overline{x}_i)$ und $x_i \in R_M$ $(i=0,1,\ldots,n)$; ferner

$$K_0 := \{x = (x_0,\ldots,x_n)' \in R_M^{n+1} : \|x-x_1\| \le l_0\} .$$

$\hat{T}''_{K_0}$ sei eine Intervallauswertung der zweiten Fréchetableitung auf K_0 und $\hat{l}_0 \in R_M$ die kleinste berechenbare obere Schranke für $\|\hat{\omega}_0 \otimes \hat{T}''_{K_0}\|$. Mit $L_0 := [l_0,l_0,l_0]$ und $\hat{L}_0 := [\hat{l}_0,\hat{l}_0,\hat{l}_0]$ gelte für die Triplexzahl $L := L_0 \otimes \hat{L}_0$ $\overline{L} \le \frac{1}{2}$. Ist $x^* = (x_0^*,\ldots,x_n^*)$ die exakte Lösung von $Tx = 0$ in K_0, dann gilt die Fehlerabschätzung

$$\|x^*-x_1\| \le \overline{F} ,$$

wobei $F = [\underline{f},f,\overline{f}] = [r,r,r] \oplus [\overline{fL},\overline{fL},\overline{fL}]$ und

$$FL = [\underline{fl},fl,\overline{fL}] = \hat{L}_0 \otimes ([1,1,1] \ominus L) \otimes L_0^2 .$$

Beweis:

Theorem 4 in ROKNE, LANCASTER [1968].

Praktisch läßt sich die Triplexarithmetik folgendermaßen beschreiben. Bei jeder auf der Maschine auszuführenden Rechenoperation werden drei Zahlen mitgeführt: Der von der Maschine durch die vorhandene Gleitkommaarithmetik gelieferte Wert

sowie zwei Maschinenzahlen, die eine sichere (natürlich möglichst gute) untere bzw. obere Schranke dieses Wertes darstellen. Hierzu ist ein umfangreicher Apparat erforderlich, da man nicht nur die Triplexgrundoperationen, sondern auch Intervallauswertungen aller benötigten Standardfunktionen, wie z.B. SQRT, SIN, COS, EXP, LN, ARCTAN, zur Verfügung stellen muß. An der von uns verwendeten Rechenanlage TELEFUNKEN TR 440 des Leibniz-Rechenzentrums der Bayerischen Akademie der Wissenschaften, München, sind die Triplexgrundoperationen als Codeprozeduren vorhanden. Die Auswertung der Standardfunktionen haben wir unter Verwendung einer Arbeit von HERZBERGER [1970] in der Programmiersprache ALGOL 60 geschrieben. Das numerische Hauptproblem bei einem Newton-Iterationsschritt ist die Auflösung eines linearen Gleichungssystems. Eine intervallarithmetische Lösung dieses Problems geben z.B. ALEFELD, HERZBERGER [1970]. Das dort beschriebene Einzelschrittverfahren diente uns als Grundlage. Auf die einzelnen Algorithmen und das ziemlich umfangreiche Programm wollen wir hier nicht näher eingehen; zur Organisation der Newton-Iteration vergleiche etwa RALL [1969], S. 153 ff. Wir wollen lediglich noch erwähnen, daß man natürlich nicht jeden Iterationsschritt als Intervallschritt ausführen muß. Vielmehr führt man soviel Iterationsschritte in Gleitkommaarithmetik aus, wie es eine herkömmliche a-priori-Schranke aufgrund der vorgegebenen Genauigkeit verlangt. Mit dem letzten berechneten m-ten Näherungsvektor $x_m = (x_{m0},\ldots,x_{mn})' \in R_M^{n+1}$ bildet man den Triplexnäherungsvektor $\mathfrak{x}_0$ gemäß

$$\mathfrak{x}_0 = (X_{00},\ldots,X_{0n})'$$

mit $$X_{0i} = [x_{mi},x_{mi},x_{mi}] \quad (i=0,\ldots,n) .$$

Bei vernünftiger Genauigkeitsforderung und nicht zu schlechten Operatorauswertungen sind dann die Voraussetzungen von Satz (10.1) erfüllbar, und man erhält die angegebene Fehlerabschätzung.

§ 11. Numerische Beispiele

1. Wir betrachten eine von BALUEV [1958] mit gewöhnlicher Kollokation und Gauss'schen Knoten numerisch behandelte Integralgleichung:

$$x(s) - 0.4 \int_0^1 e^{st} x^2(t)\,dt = s + 0.4\,\frac{2-(s^2-2s+2)e^s}{s^3}\,, \quad 0 \le s \le 1\,.$$

Die exakte Lösung dieser Gleichung ist $x^*(s) = s$.

Hat man keine Information über die Lage einer gesuchten Lösung, geht man bei Hammersteinschen Gleichungen oft mit Erfolg so vor, daß man die rechte Seite als Ausgangsnäherung wählt (z.B. MOORE [1964]). Dieser Weg wurde hier eingeschlagen.

Die Tabellierung der mit unserem Verfahren errechneten Näherung $\tilde{x}(s)$ in $[0,1]$ mit Schrittweite 0.2 liefert für $n = 8$:

s	$\tilde{x}(s)$
0	[-0.0000003 , -0.0000000 , 0.0000003]
0.2	[0.1999999 , 0.2000000 , 0.2000001]
0.4	[0.3999999 , 0.4000000 , 0.4000001]
0.6	[0.5999999 , 0.6000000 , 0.6000001]
0.8	[0.7999999 , 0.8000000 , 0.8000001]
1.0	[0.9999996 , 1.0000000 , 1.0000004]

Der Diskretisierungsfehler wurde durch 20×10^{-9} abgeschätzt, der Rundefehler durch 1.45×10^{-9}.

Der Gesamtfehler ist bei diesem Beispiel um einen Faktor 3 schlechter als bei gewöhnlicher Kollokation. Dies liegt vor allem daran, daß wir die verallgemeinerten Gewichte nur mit einer Genauigkeit von 10^{-10} berechnet haben, um die Rechenzeit vergleichbar zu halten.

Die Rechenzeit mit unserem Verfahren betrug 35.17 sec, wovon 10.45 sec auf die Berechnung der verallgemeinerten Gewichte entfielen. Die Newton-Iteration erforderte für 5 Schritte 0.42 sec, der Rechenzeitrest von 24.30 sec wurde für die Intervallrechnung benötigt.

Hingegen betrug die Rechenzeit mit gewöhnlicher Kollokation 24.70 sec, wovon lediglich 0.10 sec für die Rechnung der Gewichte erforderlich waren.

Für n = 10 ergab sich der Diskretisierungsfehler $DF = 4.2 \times 10^{-9}$ und der Rundefehler 1.48×10^{-9}. Es ist also nicht sinnvoll, für höhere Werte von n zu rechnen.

Für kleinere Werte von n ist unsere Fehlerabschätzung in diesem Beispiel nicht anwendbar, da Voraussetzung (8.13.2) von Satz (8.13) nicht erfüllt werden kann. Jedoch zeigt ein Vergleich der beiden Verfahren, daß unsere Methode auch für kleine n wesentlich bessere Werte liefert. Wir wollen den wahren Fehler der berechneten Näherungen bei gewöhnlicher Kollokation mit GK und bei verallgemeinerter Kollokation mit VK bezeichnen.

s	GK$\times 10^{-7}$		VK$\times 10^{10}$
	n = 4	n = 16	n = 4
0	2	0	3
0.2	22	2.5	43
0.4	27	1.5	16
0.6	35	1.1	5
0.8	53	6.1	0
1.0	91	0	4
RECHEN-ZEIT(sec)	0.51	1.78	1.55

2. TALBOT [1968] hat eine Intervallmethode zur direkten Lösung von Randwertaufgaben der Form

$$y''(x) = F(x,y)$$
$$y(a) = y(b) = 0$$

entwickelt, welche zu einer Näherungslösung sichere Schranken liefert. Das dort behandelte zweite Beispiel

$$y'' = e^{y}$$
$$y(0) = y(1) = 0$$

hat genau eine Lösung

$$y^*(x) = -\ln 2 + 2\ln\left\{\frac{c}{\cos(\frac{c}{2}(x-\frac{1}{2}))}\right\}^{1},$$

wobei c die eindeutig bestimmte Lösung der transzendenten Gleichung

$$c = \sqrt{2}\cos\frac{c}{4}$$

ist. Die zu diesem Randwertproblem äquivalente Integralgleichung ist

$$x(s) + \int_0^1 L(s,t)e^{x(t)}dt = 0, \quad 0\le s\le 1 ;$$

hierbei ist $L(s,t) = \begin{cases} s(1-t) \\ t(1-s) \end{cases}$ falls $\begin{matrix} s\le t \\ t\le s \end{matrix}$ die Greensche Funktion des Differentialoperators.

In der nachstehenden Tabelle haben wir die Ergebnisse für $n = 4,6,8,10$ dargestellt, wobei hier mit GK bzw. VK wieder die wahren Fehler bezeichnet sind.

s	n = 4		n = 6		n = 8		n = 10		y*(s)
	$GK \times 10^{3}$	$VK \times 10^{7}$	$GK \times 10^{4}$	$VK \times 10^{9}$	$GK \times 10^{4}$	$VK \times 10^{11}$	$GK \times 10^{4}$	$VK \times 10^{11}$	
0	0	0	0	0	0	9	0	2	0.000
0.2	7	15	32	23	18	21	11	6	-0.073
0.4	11	1	46	9	26	22	17	10	-0.109
0.6	11	1	46	9	26	24	17	9	-0.109
0.8	7	15	32	23	18	20	11	8	-0.073
1.0	0	0	0	0	0	9	0	2	0.000
FEHLER-SCHRANKE	-	29	-	27	-	80	-	190	
RECHEN-ZEIT(sec)	7.26	7.45	14.13	14.42	25.71	26.02	44.44	44.67	

Die Rechenzeit bei TALBOT beträgt bei vergleichbarer Genauigkeit im Fall n = 10 auf einer CDC-1604 7 min, 14 sec. Die dort angegebenen Fehlerschranken liegen um etwa zwei Zehnerpotenzen über dem wahren Wert.

3. Ein von MOORE [1964] behandeltes Beispiel ist

$$x(s) - \int_0^1 st\, x^2(t)dt = \frac{3}{4}\, s, \quad 0 \le s \le 1 .$$

Man kann sich überzeugen, daß in diesem speziellen Beispiel die gewöhnliche und die verallgemeinerte Kollokation zusammenfallen. Wir wollen hier die leicht modifizierte Gleichung

$$x(s) - \int_0^1 st^2 x^2(t)dt = \frac{4}{5}\, s , \quad 0 \le s \le 1$$

betrachten. Diese Gleichung hat genau die beiden Lösungen $x_1^*(s) = s$ und $x_2^*(s) = 4s$. Mit der rechten Seite als Anfangsnäherung haben wir die erste Lösung erhalten.

Bereits für $n = 4$ lieferte unsere Abschätzung für den Diskretisierungsfehler

$$DF = 15 \times 10^{-11} ,$$

für den Rundefehler

$$RF = 70 \times 10^{-11} ,$$

also einen Gesamtfehler von 85×10^{-11}.

Eine Erhöhung von n ist nicht sinnvoll, zumal die Maschinengenauigkeit bei 3×10^{-11} liegt.

Der maximale wahre Fehler an den Argumentstellen $i \times 0.1$ $(i=0,\ldots,10)$ betrug 4.3×10^{-11}.

Die gewöhnliche Kollokation lieferte bei diesem Beispiel die Lösung mit einem Maximalfehler von 3×10^{-11}, also mit derselben Genauigkeit. Die Rechenzeit war in beiden Fällen 10 sec, wobei auf die Fehlerrechnung 9.5 sec entfielen.

4. Zum Abschluß wollen wir noch eine Randwertaufgabe betrachten, die genau zwei Lösungen besitzt:

$$y'' = e^{-y}$$
$$y(0) = 0 = y(1) .$$

Auch dieses Problem wurde von TALBOT [1968] behandelt. Die beiden Lösungen dieser Aufgabe sind gegeben durch

$$y_i^*(x) = 2\{\ln[\cosh(\tfrac{c_i}{2}(x-\tfrac{1}{2}))]-\ln[\cosh(\tfrac{c_i}{4})]\} ,$$

wobei c_i $(i=1,2)$ die beiden Lösungen von

$$c = \sqrt{2}\cosh(\tfrac{c}{4})$$

sind; es ist auf drei Dezimalen $c_1 = 1.517$ und $c_2 = 10.939$. Für c_1 erhalten wir eine nahe unterhalb der x-Achse verlaufende Lösung, für c_2 eine "große Lösung" mit einem minimalen Wert von ungefähr -4.

Die äquivalente Integralgleichung lautet

$$x(s) + \int_0^1 L(s,t)e^{-x(t)}\,dt = 0 , \quad 0 \le s \le 1 ;$$

$L(s,t)$ ist wie in Beispiel 2 definiert.

Die große Lösung ist numerisch schwer zu erhalten (TALBOT [1968]). Für Werte von $n \le 14$ läßt sich die Voraussetzung (8.13.5) von Satz (8.13) nicht erfüllen, da das Residuum nicht klein genug wird. Für größere Werte von n wachsen die Rundefehler stark. Eine Fehlerabschätzung ist daher in diesem Fall mit unserer Methode nicht möglich. Ein Vergleich mit dem wahren Fehler bei gewöhnlicher Kollokation zeigt jedoch auch hier eine deutliche Überlegenheit unseres Verfahrens. Die angegebenen Rechenzeiten beziehen sich auf die Rechnung ohne Fehlerrechnung und Intervallanalysis.

s	n = 4		n = 10		$y_2^*(s)$
	GK×10	VK×10^2	GK×10^2	VK×10^4	
0	0	0	0	0	0.000
0.2	4	3	7	31	-2.122
0.4	5	21	5	8	-3.806
0.6	5	21	5	8	-3.806
0.8	4	3	7	31	-2.122
1.0	0	0	0	0	0.000
RECHEN-ZEIT(sec)	0.46	0.46	1.22	1.47	

Stückweise lineare Interpolation der drei Punkte (0;0), (0.5;-4), (1;0) lieferte uns die ziemlich grobe aber ausreichende Anfangsnäherung $\tilde{x}(s) = 8\,x-0.5\,-4$.

Wesentlich leichter zu approximieren ist die kleine Lösung $y_1^*(s)$, die wir mit der rechten Seite der Integralgleichung als Anfangsnäherung fanden. Die Ergebnisse sind in der nachstehenden Tabelle zusammengestellt.

s	n = 4		n = 6		n = 8		n = 10		$y_1^*(s)$
	$GK \times 10^3$	$VK \times 10^6$	$GK \times 10^4$	$VK \times 10^8$	$GK \times 10^4$	$VK \times 10^{10}$	$GK \times 10^4$	$VK \times 10^{12}$	
0	0	0	0	0	0	0	0	20	0.000
0.1	5	3	24	6	14	8	9	6	-0.050
0.2	10	5	46	5	26	8	16	8	-0.089
0.3	14	8	60	5	35	1	22	18	-0.118
0.4	20	5	73	5	40	4	26	0	-0.135
0.5	18	3	76	2	43	1	27	40	-0.141
FEHLER-SCHRANKE	-	8.9	-	6.2	-	11	-	3900	
RECHEN-ZEIT(sec)	7.64	7.44	14.58	14.46	26.17	26.09	44.47	44.88	

TALBOT benötigt 7 min, 46 sec, um die Lösung mit einem maximalen wahren Fehler von 10^{-11} zu berechnen; seine Fehlerschranke liegt um zwei Zehnerpotenzen über dem wahren Fehler.

LITERATURVERZEICHNIS

ALEFELD, G.
HERZBERGER, J. [1970] : ALGOL-60 Algorithmen zur Auflösung linearer Gleichungssysteme mit Fehlererfassung.
Computing 6, 28-34 (1970)

ATKINSON, K.E. [1966] : Extensions of the Nyström method for the numerical solution of linear integral equations of the second kind.
MRC Technical Summary Report # 686, Ph.D.thesis, Univ. of Wisconsin, Madison 1966

AUBIN, J.P. [1967] : Approximations des espaces de distributions et des opérateurs différentiels.
Bull.Soc.Math.France 12, 1-139 (1967)

BALUEV, A.N. [1958] : Näherungsweise Lösung nichtlinearer Integralgleichungen [russisch].
Leningr.Gos.Univ.Uch.Zap.Ser.Mat. Nauk. 33, 28-31 (1958)

CLENSHAW, C.W. [1964] : A comparison of "best" polynomial approximations with truncated Chebyshev series.
SIAM J.Numer.Anal. 1, 26-37 (1964)

ELLIOTT, D. [1965] : Truncation errors in two Chebyshev series approximations.
Math.Comp. 19, 234-248 (1965)

EHLICH, H.
ZELLER, K. [1964] : Schwankung von Polynomen zwischen Gitterpunkten.
Math.Zeitschr. 86, 41-44 (1964)

FRASER, W.
WILSON, M.W. [1966] : Remark on the Clenshaw-Curtis quadrature scheme.
SIAM Review 8, 322-327 (1966)

HERZBERGER, J. [1970] : Intervallmäßige Auswertung der Standardfunktionen in ALGOL-60.
Computing 5, 377-384 (1970)

KANTOROWITSCH, L.W.
AKILOW, G.P. [1964] : Funktionalanalysis in normierten Räumen.
Berlin: Akademie-Verlag 1964

KOLODNER, I.I. [1964] : Equations of Hammerstein type in Hilbert spaces. J.Math.Mech. 13, 700-750 (1964)

KRASNOSEL'SKII, M.A. [1964]: Topological methods in the theory of nonlinear integral equations. Oxford-London-New York-Paris: Pergamon Press 1964

[1961]: Convex functions and Orlicz spaces. Groningen: Nordhoff Publishing 1961

LOCHER, F. [1968] : Approximationsverfahren zur Gewinnung von ableitungsfreien Schranken für Quadraturfehler. Diss. Tübingen (1968)

[1969] : Fehlerabschätzungen für das Quadraturverfahren von Clenshaw und Curtis. Computing 4, 304-315 (1969)

LOCHER, F.
ZELLER, K. [1968] : Approximationsgüte und numerische Integration. Math.Zeitschr. 104, 249-251 (1968)

MEINARDUS, G. [1964] : Approximation von Funktionen und ihre numerische Behandlung. Berlin-Heidelberg-New York: Springer-Verlag 1964

MOORE, R.H. [1964] : Newton's method and variations. Nonlinear integral equations (P.M. Anselone, ed.), pp. 65-98. Madison: University of Wisconsin Press 1964

[1968] : Approximations to nonlinear operator equations and Newton's method. Num.Math.12, 23-34 (1968)

MYSOVSKIH, I.P. [1956] : Fehlerabschätzung bei der Lösung einer Integralgleichung durch mechanische Quadratur [russisch]. Vestnik Leningrad Univ. 11, 66-72 (1956)

[1962] : Über die Methode der mechanischen Quadratur zur Lösung von Integralgleichungen [russisch]. Vestnik Leningrad Univ. 17, 78-88 (1962)

NATANSON, I.P. [1965] : Constructive function theory. New York: Frederick Ungar Publishing 1965

RALL, L.B. [1969] : Computional solution of nonlinear operator equations.
New York-London-Sidney-Toronto: John Wiley & Sons 1969

ROKNE, J.
LANCASTER, P. [1969] : Automatic errorbounds for the approximate solution of equations.
Computing 4, 294-303 (1969)

SMITH, F.J. [1965] : An algorithm for summing orthogonal polynomial series and their derivatives with applications to curve fitting and interpolation.
Math.Comp. 19, 33-36 (1965)

STUMMEL, F. [1970] : Diskrete Konvergenz linearer Operatoren I.
Math.Annal. 190, 45-92 (1970)

TALBOT, Th.D. [1968] : Guaranteed error bounds for computed solutions of nonlinear two-point boundary value problems.
MRC Technical Summary Report # 875, Univ. of Wisconsin, Madison 1968

TUCKER, T.S. [1969] : Stability of nonlinear computing schemes.
SIAM J. Numer. Anal. 6, 72-81 (1969)

VAINIKKO, G.M. [1966] : The convergence of the collocation method for nonlinear differential equations. U.S.S.R. Comput.Math.math. Phys. 6, 47-58 (1966)

[1969] : Compact approximation of operators and approximate solution of operator equations.
Sov.Math.Doklady 10, 1382-1386 (1969)

L e b e n s l a u f

Am 13. November 1942 wurde ich als ältester Sohn des Kaufmanns Karl H. Maier und seiner Ehefrau Charlotte in Singen/Baden geboren.

Dort besuchte ich vom Herbst 1948 bis Ostern 1952 die Volksschule und anschließend neun Jahre das städtische Gymnasium. Nach meinem Abitur im Februar 1961 begann ich, vom Wehrdienst zurückgestellt, das Studium der Mathematik im Sommersemester 1961 an der Universität München. Mein ganzes Studium über blieb ich in München und schloß es am 5. November 1968 mit der Diplom-Hauptprüfung in Mathematik ab. Vom 15. November 1968 bis zum 31. Oktober 1970 war ich Verwalter einer wissenschaftlichen Assistentenstelle. Seit dem 1. November 1970 arbeite ich als wissenschaftlicher Angestellter auf dem Gebiet der programmgesteuerten Rechenanlagen am Mathematischen Institut der Universität München.

Seit dem 29. Juli 1964 bin ich mit der Arzthelferin Evelyn Maier, geb. Weidemann verheiratet. Wir haben eine fünfjährige Tochter.